明 細 書

文章分類装置および方法

技術分野

[0001] 本発明は、文章分類装置および方法に関し、特に文章の内容に応じて各文章を分類する文章分類装置および方法に関する。

背景技術

[0002] 高度情報化社会では、情報処理技術や情報通信技術の発展に伴い、電子化された膨大な量の情報を容易に入手できる環境が提供されつつある。このような環境を利用して入手した情報は、そのデータ量も膨大となるため、所望する情報を効率よくかつ正確に把握する必要がある。情報の内容を解析する技術として、各情報を構成する文章の内容に応じて各文章を分類する技術が研究されている。

[0003] 従来、文章を分類する技術として、予め各分類の内容を示すラベルを用意し、各文章の内容を所定のアルゴリズムで解析し、用意した各ラベルごとにそれぞれの文章を分類するものが提案されている(例えば、永田昌明他,「テキスト分類ー学習論理の見本市ー」,情報処理,42巻1号,2001年1月など参照)。このような技術は、文章の分類に際し、予め各分類の内容を示すラベルを用意し、各種の学習アルゴリズムを用いて、これらラベルを各文章に対して精度よく割り当てることにより、各文章をラベルごとに分類しようとするものである。

発明の開示

発明が解決しようとする課題

[0004] しかしながら、このような従来の文章分類技術では、予めラベルを用意する必要があるため、分類対象となる各文章の内容をある程度把握して適切なラベルを選択して設定しておく必要がある。したがって、このラベル選択に際し、文章量が多くその内容が広範囲にわたる場合には大きな作業負担を要するという問題点があった。また、分類に用いるラベルは主観的に選択されることから、得られる分類そのものが限定的となり、想定しうる範囲を超えた新たな観点さらにはより大きな観点から文章を分類できないという問題点があった。

本発明は、以上のような問題点を解消するためになされたものであり、比較的少ない作業負担で、主観にとらわれることなく柔軟に分類できる文章分類装置および方法を提供することを目的としている。

課題を解決するための手段

[0005] 本発明にかかる文章分類装置は、1つ以上の単語からなるタームを複数有するタームリストと、文章集合に含まれる各文章と各タームとの関係を2次元表現したDTマトリクスを生成するDTマトリクス生成手段と、グラフ理論で用いられるDM分解法に基づいてDTマトリクス生成手段で得られたDTマトリクスを変形することにより、関連する文章のブロックからなるクラスタを有する変形DTマトリクスを生成するDTマトリクス変形手段と、このDTマトリクス変形手段で得られた変形DTマトリクス上の各クラスタと、これらクラスタで分類される各文章との関係に基づき、文章集合に関する分類を生成する分類生成手段とを備える。

この際、上記分類生成手段として、DTマトリクス変形手段で得られた変形DTマトリクス上のクラスタごとに、当該クラスタに属する文章を同一分類として出力する文章分類手段を備えてもよい。

あるいは、上記分類生成手段として、変形DTマトリクス上のクラスタごとに、当該クラスタに属する各文章のタームから仮想代表文章を生成する仮想代表文章生成手段と、DTマトリクス生成手段で生成したDTマトリクスを初期状態として、DTマトリクス変形手段でDTマトリクスから生成された変形DTマトリクス上のクラスタごとに仮想代表文章生成手段で仮想代表文章を生成し、仮想代表文章を当該変形DTマトリクスに追加するとともに仮想代表文章のクラスタに属する文章を当該変形DTマトリクスから削除して次のクラスタリング処理に用いる新たなDTマトリクスを生成し、クラスタごとに少なくとも当該クラスタを構成する文章に関する情報を大分類データとして出力する、というクラスタリング処理を繰り返し行うことにより文章の大分類を生成する大分類生成手段とを備えてもよい。

[0006] また、本発明にかかる文章分類方法は、文章集合に含まれる各文章と1つ以上の単語からなるタームを複数有するタームリストの各タームとの関係を2次元表現したDTマトリクスを生成するDTマトリクス生成ステップと、グラフ理論で用いられるDM分解

法に基づいてDTマトリクスを変形することにより、関連する文章のブロックからなるクラスタを有するクラスタを有する変形DTマトリクスを生成するDTマトリクス変形ステップと、変形DTマトリクス上の各クラスタと、これらクラスタで分類される各文章との関係に基づき、文章集合に関する分類を生成する分類生成ステップとを備える。

この際、上記分類生成ステップとして、変形DTマトリクス上のクラスタごとに、当該クラスタに属する文章を同一分類として出力する文章分類ステップを備えてもよい。

あるいは、上記分類生成ステップとして、変形DTマトリクス上のクラスタごとに、当該クラスタに属する各文章のタームから仮想代表文章を生成する仮想代表文章生成ステップと、DTマトリクス生成ステップで生成したDTマトリクスを初期状態として、DTマトリクス変形ステップでDTマトリクスから生成された変形DTマトリクス上のクラスタごとに仮想代表文章生成ステップで仮想代表文章を生成するステップと、仮想代表文章を当該変形DTマトリクスに追加するとともに仮想代表文章のクラスタに属する文章を当該変形DTマトリクスから削除して次のクラスタリング処理に用いる新たなDTマトリクスを生成するステップと、クラスタごとに少なくとも当該クラスタを構成する文章に関する情報を大分類データとして出力するステップと、からなるクラスタリング処理を繰り返し行うことにより文章の大分類を生成する大分類生成ステップとを備えてもよい。

発明の効果

[0007] 本発明によれば、文章集合内の各文章とタームリスト内の各タームとから生成されたDTマトリクスがDM分解されて、得られた変形DTマトリクス上の各クラスタごとに、当該クラスタに属する各文章が1つの分類として抽出されるため、各分類に対応したラベルを予め用意することなく各文章を分類できる。これにより、従来のように分類対象となる各文章の内容をある程度把握して適切なラベルを選択する必要がなくなることから、出現頻度など分類に直接関係のない尺度で選択した単語からタームを構成することができ、ラベル選択のための作業負担を大幅に軽減できる。したがって、比較的少ない作業負担で、ラベルなどの予め用意された主観にとらわれることなく柔軟に分類できる。

また、DTマトリクスに対する変形処理により新たなクラスタを生成するとともに、そのクラスタをその仮想代表文章で置換することにより新たなDTマトリクスを生成するクラ

スタリング処理を繰り返し実行するようにしたので、従来のようにラベルを用意することなく、新たなDTマトリクスから順次新たなクラスタ、すなわちクラスタを含むより大きなクラスタすなわち大分類を得ることができ、想定しうる範囲を超えたより大きな観点から文章集合内の各文章を分類できる。したがって、比較的少ない作業負担で、ラベルなどの予め用意された主観にとらわれることなく柔軟に分類できる。

図面の簡単な説明

[0008] [図1]図1は、本発明の一実施の形態にかかる文章分類装置の構成を示すブロック図である。

[図2]図2は、DTマトリクス生成処理を示すフローチャートである。

[図3]図3は、文章集合の構成例である。

[図4]図4は、タームリストの構成例である。

[図5]図5は、DTマトリクスの構成例である。

[図6]図6は、DM分解処理を示すフローチャートである。

[図7A]図7Aは、DM分解処理の過程を示す2部グラフである。

[図7B]図7Bは、DM分解処理の過程を示す2部グラフである。

[図7C]図7Cは、DM分解処理の過程を示す2部グラフである。

[図7D]図7Dは、DM分解処理の過程を示す2部グラフである。

[図7E]図7Eは、DM分解処理の過程を示す2部グラフである。

[図7F]図7Fは、DM分解処理の過程を示す2部グラフである。

[図8A]図8Aは、DTマトリクスの例である。

[図8B]図8Bは、変形DTマトリクスの例である。

[図9]図9は、文章分類処理を示すフローチャートである。

[図10]図10は、文章分類処理を示す説明図である。

[図11]図11は、ラベル生成処理を示すフローチャートである。

[図12]図12は、ラベル生成処理を示す説明図である。

[図13]図13は、文章編成処理を示すフローチャートである。

[図14]図14は、文章編成処理を示す説明図である。

[図15]図15は、要約作成処理を示すフローチャートである。

[図16]図16は、要約作成処理を示す説明図である。

[図17]図17は、指標生成処理を示すフローチャートである。

[図18]図18は、大分類生成処理を示すフローチャートである。

[図19]図19は、大分類生成処理の実行例を示す説明図である。

[図20]図20は、大分類ラベル生成処理を示すフローチャートである。

[図21]図21は、DTマトリクスの生成例(初期状態)である。

[図22]図22は、DTマトリクスの生成例(最終ステップ)である。

発明を実施するための最良の形態

[0009] 次に、本発明の実施の形態について図面を参照して説明する。

[第1の実施の形態]

まず、図1を参照して、本発明の第1の実施の形態にかかる文章分類装置について説明する。図1は本発明の第1の実施の形態にかかる文章分類装置の構成を示すブロック図である。この文章分類装置1は、全体としてコンピュータからなり、制御部10、記憶部20、操作入力部30、画面表示部40、およびデータ入出力インターフェース部(以下、データ入出力I/F部という)50が設けられている。

[0010] 制御部10は、CPUなどのマイクロプロセッサとその周辺回路からなり、記憶部20に予め格納されているプログラム(図示せず)を実行して、上記ハードウェアとプログラムとを協働させることにより、文章分類処理のための各種機能手段を実現する。記憶部20は、ハードディスクやメモリなどの記憶装置からなり、制御部10での処理に用いる各種情報を格納する。これら情報としては、分類対象となる各文章からなる文章集合21、各文章の内容を把握するための複数の重要語からなるタームリスト22、さらには文章を大分類した結果を示す大分類データ23が記憶されている。

[0011] 操作入力部30は、キーボードやマウスなどの入力装置からなり、利用者の操作を検出して制御部10へ出力する。画面表示部40は、CRTやLCDなどの画面表示装置からなり、制御部10での処理内容や処理結果を表示出力する。データ入出力I/F部50は、外部装置(図示せず)や通信ネットワーク(図示せず)と接続するための回路部であり、文章集合21、タームリスト22、大分類データ23のほか、得られた処理結果や制御部10で実行するプログラムをやり取りする際に用いられる。

[0012] 制御部10には、機能手段として、DTマトリクス生成手段11、DTマトリクス変形手段12、文章分類手段(分類生成手段)13、ラベル生成手段14、文章編成手段15、要約作成手段16、タームリスト編集手段17、タームリスト生成手段18、指標生成手段19、大分類生成手段(分類生成手段)71、仮想代表生成手段(分類生成手段)72、および大分類ラベル生成手段73が設けられている。

[0013] 本実施の形態において、DTマトリクスとは、各文章Dと各タームTとの関係を2次元的に表現した行列を指す。この際、上記関係は、文章D中におけるタームTの存在有無からなり、文章Dとターム Tとをそれぞれマトリクスの列と行に対応させ、ある文章Diがあるターム Tjを含む場合には、DTマトリクスのj, i成分を「1」とし、含まない場合には「0」とすることにより、文章Dとターム Tの関係を表している。さらに、このDTマトリクスを2部グラフの一表現形態と見なし、2部グラフのグラフ理論で用いられるDM分解法に基づきDTマトリクスを変形し、得られた変形DTマトリクス上に現れるクラスタに基づき、各文章Dを分類するようにしたものである。

[0014] DTマトリクス生成手段11は、分類対象となる各文章D(Document)とタームリスト22を構成する各ターム T(Term)とからDT(Document-Term)マトリクスを生成する機能手段である。DTマトリクス変形手段12は、DTマトリクス生成手段11で生成されたDTマトリクスをDM(Dulumage-Mendelsohn)分解法に基づき変形する機能手段である。DM分解法とは、具体的には、DTマトリクスに対し、行操作(行同士を入れ替える操作)または列操作(列同士を入れ替える操作)を施して、三角行列化する処理である。この三角行列化されたDTマトリクスを変形DTマトリクスと呼ぶ。

[0015] 文章分類手段13は、DTマトリクス変形手段12で得られた変形DTマトリクス上に現れるブロック化されたクラスタに基づき、文章集合21の各文章を分類する機能手段である。ラベル生成手段14は、各クラスタごとに、当該クラスタに属する各文章Dと強連結の関係にあるターム Tを、当該クラスタのラベルとして出力する機能手段である。文章編成手段15は、変形DTマトリクスにおける文章Dの並び順に基づき、文章集合21の各文章を並び替えて出力する機能手段である。要約作成手段16は、文章Dと強連結の関係にあるターム Tを含む文を、当該文章Dの要約として出力する機能手段である。

[0016] タームリスト編集手段17は、操作入力部30からの操作に応じて、記憶部20のタームリスト22に対するタームTの追加／削除を行う機能手段である。タームリスト生成手段18は、記憶部20の文章集合21に含まれる各文章Dを解析して、各文章Dの特徴を効果的に表現する語すなわち重要語を抽出し、これら重要語からなるタームTを用いてタームリスト22を生成する機能手段である。指標生成手段19は、タームリスト編集手段17で編集されたタームリストについて、その編集前後におけるDTマトリクスに基づき当該編集による分類への影響を示す指標を生成する機能手段である。

[0017] 大分類生成手段71は、DM分解法を用いたDTマトリクス変形手段12でのDTマトリクス変形処理をクラスタリング処理として繰り返し実行し、各クラスタリング処理で得られた変形DTマトリクスから得られたクラスタに基づき、文章集合21の各文章の大分類を生成する機能手段である。仮想代表生成手段72は、大分類生成時に、変形DTマトリクスから得られたクラスタから、そのクラスタに含まれる文章を仮想的に代表する仮想代表文章を生成する機能手段である。大分類ラベル生成手段73は、大分類生成手段71で生成された各クラスタすなわち大分類のラベルを生成する機能手段である。なお、大分類生成手段71、仮想代表生成手段72、および大分類ラベル生成手段73は、後述する第2の実施の形態で用いられる。

[0018] [第1の実施の形態の動作]

次に、図2を参照して、本発明の第1の実施の形態にかかる文章分類装置の動作について説明する。図2は本発明の第1の実施の形態にかかる文章分類装置のDTマトリクス生成処理を示すフローチャートである。制御部10は、操作入力部30からの指示に応じて、文章分類処理に用いるDTマトリクスを生成するため、図2のDTマトリクス生成処理を開始する。まず、DTマトリクス生成手段11は、記憶部20に格納されている文章集合21を読み込むとともに(ステップ100)、タームリスト22を読み込む(ステップ101)。

[0019] 図3に文章集合21の構成例を示す。この例は、「ストレス」についてWeb上で多数の回答者に自由に文章を記述してもらったものを集計したものであり、各文章Dごとに当該文章Dを管理するための文章番号Diとその文章を記述した回答者の識別情報とが割り当てられている。図4はタームリスト22の構成例である。このタームリスト22

は、所定のアルゴリズムに基づき各文章Dを解析し、得られた重要語の種別とその前後関係とから各タームTを構成したものであり、各タームTごとに当該タームTを管理するターム番号Tjが割り当てられている。

[0020] 各タームTは、2つの重要語のうち、前方に位置するキーワード前と後方に位置するキーワード後からなり、それぞれのキーワードごとにそのキーワードの内容を示す単語とその単語の品詞属性種別とが規定されている。また、各タームTには、後述するタームリスト生成処理により文章集合21から算出された、文章分類に用いる上での重みを示す重要度が対応付けられている。例えばターム「1」は、「ストレス」と「解消」という2つのキーワードからなり、その位置関係は「ストレス」が前方に位置するものと規定されている。

DTマトリクス生成手段11は、文章集合21内の各文章について、あるしきい値以上の重要度を持ったタームリスト22の各タームTが存在するか否かチェックし、その結果からDTマトリクスを生成する(ステップ102)。図5にDTマトリクスの構成例を示す。このDTマトリクス11Aは、行方向(縦方向)にタームTが並べられており、列方向(横方向)に文章Dが並べられている。そして、各文章Dとタームの交差位置に、当該文章Dにおけるタームの存在有無が2進数で記載されている。ここでは、文章Dにタームが存在する場合は「1」が設定され、存在しない場合は「0」が設定されている。したがって、この例によれば、例えば文章D1には、タームT4, T7が含まれていることがわかる。またタームT2は、文章D2, D4に含まれていることがわかる。

[0021] 続いて、DTマトリクス変形手段12は、このようにしてDTマトリクス生成手段11で生成されたDTマトリクス11Aを、DM分解法に基づき変形して変形DTマトリクス11Bを生成し(ステップ103)、これを記憶部20に格納して、一連のマトリクス生成処理を終了する。一般に、グラフ理論では、2つの集合に属するそれぞれの点とこれら点を結ぶ辺とからなる2部グラフを、各点間の関連性に基づき分離する手法として、DM分解法が用いられる。本実施の形態では、DTマトリクス11Aを、文章Dからタームへの辺により結びつけられた2部グラフの一表現形態と見なすことができることに着目し、グラフ理論におけるDM分解法をDTマトリクス11Aに適用し、得られた変形DTマトリクスに基づき文章Dを分類するようにしたものである。

[0022] [DM分解処理]

ここで、図6および図7A一図7Fを参照して、2部グラフにおけるDM分解処理について説明する。図6はDM分解処理を示すフローチャートである。図7A一図7FはDM分解処理の過程を示す2部グラフである。以下では、文章DおよびタームTからなる2つの点集合と、これら点を結ぶ辺からなる2部グラフGを処理対象とし、これをDM分解法により複数のグラフに分離する場合を例として説明する。なお、これら処理では、制御部10内部のメモリまたは記憶部20から各種データを読み出して、制御部10で所定の演算を行い、その結果を再び記憶するという動作が繰り返し行われる。

[0023] まず、図7Aに示すように、処理対象となる2部グラフGの各辺について、文章Dからタームへの有向辺を生成する(ステップ200)。そして、図7Bに示すように、文章D側に点sを用意し、点sから文章Dの各点に対して有向辺を生成する(ステップ201)。同様にして、タームT側に点tを用意し、タームTの各点から点tに対して有向辺を生成する(ステップ202)。

[0024] 次に、これら辺を介して点sから点tへ向かう経路を検索する(ステップ203)。例えば図7Bでは、辺250, 251, 252からなる経路を介して点sから点tへ向かうことができる。このような経路が存在する場合は(ステップ203:YES)、当該経路を構成する各辺を削除するとともに(ステップ204)、当該経路上の文章DからタームTへの有向辺とは逆向きの有向辺を、初期状態で空の2部グラフである最大マッチングMに生成し(ステップ205)、ステップ203へ戻って次の経路を検索する。図7Cでは、有向辺251に対応する逆向きの有向辺253が最大マッチングMに生成されている。ステップ203において、すべての経路の検索が終了して新たな経路が検索されなかった場合(ステップ203:NO)、最大マッチングMが完成したことになる。

[0025] このようにして、図7Dに示すような最大マッチングMを完成させた後、最大マッチングMに属する各有向辺254を処理対象Gへ含める(ステップ206)。これにより、図7Eに示すように、処理対象Gにおいて、最大マッチングMとして選択された辺255については、文章DからタームTへの有向辺とその逆方向の有向辺とから構成されることになる。

[0026] 次に、タームTの各点のうち最大マッチングMに用いられなかった点、例えば自由

点256を選択し(ステップ207)、図7Fに示すように、処理対象Gの各辺を介して当該自由点256に到達可能な点の集合をクラスタ260とする(ステップ208)。同様にして、文章Dの各点のうち最大マッチングMに用いられなかった点、例えば自由点257を選択し(ステップ209)、処理対象Gの各辺を介して当該自由点257に到達可能な点の集合をクラスタ262とする(ステップ210)。そして、残りの文章DおよびタームTの各点のうち、双方向に到達可能な経路を有する点集合すなわち強連結をなす点集合をクラスタ261とし(ステップ211)、一連のDM分解処理を終了する。このようにして、公知のDM分解法では、各クラスタが所定の順序で生成され、三角行列化された変形DTマトリクスが得られる。

[0027] 制御部10では、以上のようにして、図2のDTマトリクス生成処理を実行することにより、DTマトリクス生成手段11で文章集合21とタームリスト22とからDTマトリクス11Aを生成するとともに、DTマトリクス変形手段12でDTマトリクスに対して図6のDM分解処理を適用することにより、各文章Dがクラスタごとに分離された変形DTマトリクス11Bを生成する。

[0028] 図8AにDTマトリクス11Aの例を示す。また図8Bに変形DTマトリクス11Bの例を示す。ここでは、各文章Di内においてタームTjが存在する場合、列方向(横方向)に配置された文章Diと行方向(縦方向)に配置されたタームTiとの交点にドットが配置されており、タームTjが存在しない場合は空白となっている。図8AのDTマトリクス11Aでは、ドットがランダムに分布しているが、図8Bの変形DTマトリクス11Bでは、ドットが断片的ではあるが斜め方向に連続して密集しており、この部分270にクラスタが並んでいることがわかる。また、変形DTマトリクス11Bでは、左下側にドットが存在せず、右上側にドットが多く存在しており、上三角行列化されていることがわかる。

[0029] [文章分類処理]

文章分類装置1の制御部10では、文章集合21を分類する場合、まず前述のDTマトリクス生成処理(図2参照)を実行した後、図9の文章分類処理を実行する。図9は文章分類処理を示すフローチャートである。まず、文章分類手段13は、DTマトリクス変形手段12で生成した変形DTマトリクス11B上にブロック化されて現れた各クラスタを識別する(ステップ110)。この際、各クラスタについては、変形DTマトリクス11Bを

生成した際に分離した部分グラフに基づき識別してもよく、変形DTマトリクス11B上のデータ(ドット)の並びから識別してもよい。

[0030] 図10に文章分類処理の説明図を示す。この例では、変形DTマトリクス11B上にクラスタ60が存在している。このクラスタ60は、2部グラフで表現した場合の部分グラフ61をなしており、他の文章やタームと関連性が小さい。なお、クラスタ境界が明確な完全グラフをなす場合もある。変形DTマトリクス11Bでは、列方向(横方向)に文章Dが並んでおり、クラスタ60の列方向に並ぶ文章DすなわちD363, D155, D157, D5, D13, D8が、このクラスタ60に属する文章Dとなる。文章分類手段13は、識別された各クラスタに属する各文章からなる部分集合62を1つの分類として、文章集合21から抽出して分類し(ステップ111)、その結果を例えば画面表示部40で表示出力し、あるいは記憶部20へ格納して、一連の文章分類処理を終了する。

[0031] このように、本実施の形態では、変形DTマトリクス11B上でブロック化されたクラスタごとに、当該クラスタに属する各文章を1つの分類として抽出出力するようにしたので、各分類に対応したラベルを予め用意することなく各文章を分類できる。したがって、従来のように分類対象となる各文章の内容をある程度把握して適切なラベルを選択する必要がなくなることから、出現頻度など分類に直接関係のない尺度で選択した単語からタームを構成することができ、ラベル選択のための作業負担を大幅に軽減できる。

[0032] また、これらクラスタは、複数のタームを橋渡しとして関連付けられた複数の文章から構成されているため、同一タームを含む文章を1つの分類として抽出することができるだけでなく、これら文章内にほぼ共通して存在する他のタームについても、そのタームを含む文章を同一分類として抽出でき、内容に共通性や関連性を持つ文章を1つの分類として容易に抽出できる。したがって、従来のように予め用意したラベルの有無のみに基づき文章を分類する場合と比較して、そのラベルに限定された主観的な分類ではなく、想定しうる範囲を超えた新たな観点から文章の内容や話題に沿って柔軟に分類を行うことができる。

[0033] [ラベル生成処理]

文章分類装置1の制御部10では、文章分類手段13で分類された各文章の分類ご

とにラベルを生成する場合、まず前述のDTマトリクス生成処理(図2参照)および文章分類処理(図9参照)を実行した後、図11のラベル生成処理を実行する。図11はラベル生成処理を示すフローチャートである。まず、ラベル生成手段14は、ラベルを生成する対象となる分類すなわちクラスタに属する各文章Dについて、これら文章Dと強連結の関係にあるタームTを変形DTマトリクス11Bから選択する(ステップ120)。

[0034] 図12にラベル生成処理の説明図を示す。この例では、任意の分類に属する文章を示す部分集合62について、各文章Dと強連結の関係にあるタームT(63)がそれぞれ選択されている。なお、強連結とは、変形DTマトリクス11Bで各文章Dをクラスタごとに分類した際、その2部グラフにおいて、文章Dとタームtとが互いに双方向の辺で結ばれたペアをいう。通常、これら強連結をなす文章Dとタームtとは、変形マトリクス上の当該クラスタにおいて対角線上に並ぶ。次に、選択した各タームTの単語を当該分類のラベル64として出力し(ステップ121)、その結果を例えば画面表示部40で表示出力し、あるいは記憶部20へ格納して、一連のラベル生成処理を終了する。

[0035] このように、本実施の形態では、対象となる分類のクラスタに属する各文章と強連結の関係にあるタームTを、当該分類のラベルとして出力するようにしたので、本実施の形態のように予め用意されたラベルに基づき文章を分類するものではない場合でも、各分類の特徴を単語で表現した適切なラベルを容易に生成できる。

[0036] [文章編成処理]

文章分類装置1の制御部10では、各文章Dの並びを編成する場合、まず前述のDTマトリクス生成処理(図2参照)を実行した後、図13の文章編成処理を実行する。図13は文章編成処理を示すフローチャートである。まず、文章編成手段15は、変形DTマトリクス11B上での並びに基づき、各文章Dを並び替える(ステップ130)。図14に文章編成処理の説明図を示す。前述したように、DTマトリクスをDM分解法により変形して得られた変形DTマトリクス11Bにおいて、各文章Dはタームtを仲立ちとして互いに関連性の高いものが隣接して並んでいる。文章編成手段15は、このような変形DTマトリクス11Bに基づき並び変えられた文章Dを編成し、編成された各文章65を出力し(ステップ131)、その結果を例えば画面表示部40で表示出力し、あるい

は記憶部20へ格納して、一連の文章編成処理を終了する。

[0037] 特に、変形DTマトリクス11Bには、文章Dおよびターム Tの並びに所定の半順序が存在する。例えば、DTマトリクス11Aは、タームTを変数とする文章Dの線形連立方程式を示す行列と見なすことができ、変形DTマトリクス11Bは、これら各方程式の解Gが求まる順序にほぼ沿った順序で文章Dが並び替えられた結果を示している。このことからも、変形DTマトリクス11B上の文章Dの並びには、前後の文章Dとの関連性が高いことがわかる。

[0038] このように、本実施の形態では、変形DTマトリクス上の文章Dの並びに基づき、各文章Dを並び替えて出力するようにしたので、共通のタームすなわち単語を持った関連性の高い文章が順に得られることになり、前後の文章Dと話題の共通性が得られる。したがって、内容が類似した文章が前後に並べられていることから、アトランダムに文章Dを読む場合と比較して、文脈が途切れることなく読むことができクラスタさらには文章集合全体の内容を容易に把握できる。この際、任意のクラスタすなわち分類に含まれる各文章Dを文章編成の対象として1つの文章を生成してもよく、文章集合21に含まれるすべての文章Dを文章編成の対象として1つの文章を生成してもよい。

[0039] [要約作成処理]

文章分類装置1の制御部10では、複数の文からなる任意の文章Dの要約を作成する場合、前述のDTマトリクス生成処理(図2参照)を実行した後、図15の要約作成処理を実行する。図15は要約作成処理を示すフローチャートである。まず、要約作成手段16は、対象となる文章Dについて、前述したラベル生成処理と同様にして、その文章Dと強連結の関係にあるタームTを変形DTマトリクス11Bから選択する(ステップ140)。

[0040] 図16に要約作成処理の説明図を示す。通常、文章D(66)は、複数の文から構成されており、これら文のいずれかに文章Dと強連結のタームT(67)が含まれていることになる。この際、このタームTは文章Dの特徴を示していることになる。要約作成手段16は、このタームTを含む文を当該文章Dから選択して、これら文を当該文章Dの要約68として出力し(ステップ141)、その結果を例えば画面表示部40で表示出力し、あるいは記憶部20へ格納して、一連の要約作成処理を終了する。

[0041] このように、本実施の形態によれば、対象となる文章Dと強連結の関係にあるタームTに基づいて、そのタームを含む文を当該文章Dの要約として出力するようにしたので、文章Dの要約を極めて容易にかつ適切に作成できる。

[0042] [タームリスト生成処理]

タームリスト生成手段18は、文章集合21からタームリスト22を自動生成するものである。文章からその文章を特徴付ける重要語を抽出する方法として、各種のアルゴリズムが提案されている。例えば、各単語の重要度を算出し、その重要度に基づき重要語を選択するTFIDF(Term Frequency Inverse Document Frequency)などのアルゴリズムを用いてもよい。あるいは、言語学的な解釈に基づかないフレーズ(共起語)を、辞書を用いることなく抽出するKeyGraphというアルゴリズムを用いてもよい(例えば、北研二他,「情報検索アルゴリズム」,共立出版,2002年など参照)。

[0043] タームリスト生成手段18では、このような公知のアルゴリズムを用いてタームリスト22を生成する。本実施の形態では、これら単語を特定するため、各単語の品詞属性を形態素解析により予め求めておき、単語のとの品詞属性をペアとして重要語を構成している。また、本実施の形態では、2つの重要語の出現順序を規定したものをタームとして定義しており、これにより文章の内容をより適切にタームで表現可能となっている。なお、このタームリスト22については、タームリスト編集手段17で、操作入力部30からの指示に基づき生成してもよく、データ入出力I/F部50を介して予め用意されたものを装置外部から入力するようにしてもよい。

[0044] [指標生成処理]

タームリスト22は、変形DTマトリクス11Bを生成して文章を分類する上で重要なファクタとなることから、タームリスト編集手段17で、このタームリストを編集可能としている。本実施の形態では、編集されたタームリストについて、制御部10の指標生成手段19により客観的な評価値を算出し、その編集に対する指標を生成する。以下、図17を参照して、指標生成手段19における指標生成処理について説明する。図17は指標生成処理を示すフローチャートである。

[0045] まず、タームリスト編集手段17により、タームリスト22についてタームTkを追加または削除し、新たなタームリストが生成されたものとする(ステップ150)。指標生成手段

19では、編集前後のタームリストのそれぞれについて、DTマトリクス生成手段11によりDTマトリクスを生成し(ステップ151)、各DTマトリクスごとに平均文章類似度Qを算出する(ステップ152)。平均文章類似度Qは、2つの文章Di, Dj間の類似度sim(Di, Dj)をすべての文章間について算出し平均したものであり、文章Dの数をNとした場合、Qは次の数1で算出される。

[0046] [数1]

$$Q = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{\substack{j=1 \\ i \neq j}}^{N} sim(Di, Dj) \quad \cdots(1)$$

[0047] この際、類似度sim(Di, Dj)は、当該変形DTマトリクスに基づき、文章Di, Djにおける各タームTの有無を0／1で示すベクトルをX, Yとした場合、例えば数2ー数4により算出される。特に、数2はベクトルX, Yの内積を類似度とするもの、数3はベクトルX, YのDice係数を類似度とするもの、数4はベクトルX, YのJaccard係数を類似度とするものである。

[0048] [数2]

$$sim(Di, Dj) = |X \cap Y| = \sum_{i=1}^{t} xi \cdot yi \quad \cdots(2)$$

[0049] [数3]

$$sim(Di, Dj) = \frac{2|X \cap Y|}{|X| + |Y|} \quad \cdots(3)$$

[0050] [数4]

$$sim(Di, Dj) = \frac{|X \cap Y|}{|X| + |Y| - |X \cap Y|} \quad \cdots(4)$$

[0051] このようにして、指標生成手段19は、編集前のタームリストから生成されたDTマトリクスに基づき平均文章類似度Qを算出するとともに、編集後のタームリストから生成されたDTマトリクスに基づき平均文章類似度Qkを算出して、これらの差ΔQを数5で算出し、指標値として画面表示部40から表示出力する(ステップ153)。

[0052] [数5]

$$\Delta Q = Qk - Q \quad \cdots(5)$$

[0053] ここで、差ΔQがゼロより大きい場合は(ステップ154:YES)、編集後のタームリストから生成されたDTマトリクスのほうが、各文章の類似度が大きくなり、各文章が効果的に分類できることから、当該編集は有効である旨を画面表示部40へ表示出力し(ステップ155)、一連の指標生成処理を終了する。

[0054] また、ステップ154において、差ΔQがゼロ以下の場合は(ステップ154:NO)、編集後のタームリストから生成されたDTマトリクスのほうが、各文章の類似度が小さくなり、各文章が効果的に分類できないことから、当該編集は無効である旨を画面表示部40へ表示出力し(ステップ156)、一連の指標生成処理を終了する。なお、指標としては、ΔQだけを表示出力して作業者に編集の有効性を判断させるようにしてもよい。また当該編集に対する有効／無効だけを表示出力してもよい。

[0055] このように、本実施の形態では、指標生成手段19により、編集前後のタームリストから生成されたDTマトリクスに基づき平均文章類似度Qを算出し、その変化により当該編集の有効性を示す指標を生成するようにしたので、タームリスト22に対する編集の有効性を容易に把握することができる。したがって、容易かつ適切にタームリストを編集でき、この編集により所望の意図や目的に応じて効率よく文章を分類することができる。また、DTマトリクスから得られた平均文章類似度に基づき指標を生成するようにしたので、文章を分類する必要がなくなり指標生成に要する処理を簡素化できる。したがって、当該編集に対する有効／無効を迅速に判断でき、タームリストの編集に要する作業負担を大幅に軽減できる。

[0056] なお、平均文章類似度Qを用いて当該編集に対する有効／無効を判断する場合について説明したが、これに限定されるものではない。例えば文章を分類した結果、例えば分類数や1分類に属する文章数などに基づき当該編集に対する有効／無効を判断するようにしてもよい。

[0057] [第2の実施の形態]

次に、図18を参照して、本発明の第2の実施の形態にかかる文章分類装置について説明する。図18は、本発明の第2の実施の形態にかかる文章分類装置での大分類生成処理を示すフローチャートである。なお、本実施の形態にかかる文章分類装

置の構成については、前述した第1の実施の形態にかかる文章分類装置（図1参照）と同様であり、ここでの詳細な説明については省略する。

[0058] 前述した第1の実施の形態では、各文章と各タームとの関係を2次元表現したDTマトリクスを生成し、グラフ理論で用いられるDM分解法に基づいてそのDTマトリクスを変形し、得られた変形DTマトリクス上で識別されたクラスタを用いて、各文章を分類する処理について説明した。前述の分類処理では、各文章をクラスタごとに文章集合としてある程度分類できるものの、1つ以上のクラスタを含むより大きな分類すなわち大分類や、クラスタ間の階層的関係については対応できない。本実施の形態では、文章分類装置1の制御部10に設けた、大分類生成手段71、仮想代表生成手段72、および大分類ラベル生成手段73を用いて、各文章の大分類を生成するようにしたものである。

[0059] [第2の実施の形態の動作（大分類生成処理）]

次に、図18を参照し、本発明の第2の実施の形態にかかる文章分類装置の動作として、各文章の大分類を生成する大分類生成処理について詳細に説明する。

制御部10は、操作入力部30からの指示に応じて、大分類生成手段71により、図18の大分類生成処理を開始する。まず、大分類生成手段71は、DTマトリクス生成手段11を用いて、記憶部20に格納されている文章集合21とタームリスト22を読み込み、前述と同様のDTマトリクス生成処理を行うことにより、各文章と各タームとの関係を2次元表現したDTマトリクスを生成する（ステップ160）。

[0060] 次に、大分類生成手段71は、DTマトリクス変形手段12を用いて、グラフ理論におけるDM分解法を上記DTマトリクスに適用し、前述と同様にして各文章がクラスタごとに分離された変形DTマトリクス11Bを生成する（ステップ161）。そして、大分類生成手段71は、前述した文章分類手段13と同様にして、得られた変形DTマトリクス上でブロック化された各クラスタを識別する（ステップ162）。

[0061] ここで、新たなクラスタが識別された場合は（ステップ163：YES）、仮想代表生成手段72を用いて、新たなクラスタごとにそのクラスタを仮想的に代表する仮想代表文章を生成する。仮想代表生成手段72では、まず、新たなクラスタに属する各文章の特徴量を取得し、これら特徴量の和集合から仮想代表文章を生成する。例えば、各文

章の特徴量Kiが数6のように、1つ以上の特徴量k1ーknで表現される場合、仮想代表文章K'は、数7で得られる。

[0062] [数6]

$$Ki = \{k1, k2, \cdots, kn\} \quad \cdots(6)$$

[0063] [数7]

$$K' = K1 \cup K2 \cup \cdots \cup Km \quad \cdots(7)$$

[0064] この際、例えば特徴量として前述のようにタームを用いる場合、仮想代表文章は、新たなクラスタに属する各文章が持つタームをすべて含む和集合となり、その内容は、各タームを構成するキーワードの羅列から構成される。

[0065] 大分類生成手段71は、仮想代表生成手段72により、上記のようにして新たなクラスタごとにその仮想代表文章を生成して新たな文章番号を付与し(ステップ164)、これら仮想代表文章を他の実際の文章(実文章)と同様の文章として変形DTマトリクスへ追加するとともに(ステップ165)、新たなクラスタに属する各文章を変形DTマトリクスから削除する(ステップ166)。これにより、変形DTマトリクス上では、仮想代表文章とこれに含まれる各タームとの交点にドットが追加配置されるとともに、元の各文章に対応するドットが削除され、新たなクラスタを構成する各文章が仮想代表文章で置換された新たなDTマトリクスが生成される。

[0066] この後、大分類生成手段71は、新たなクラスタの構成、例えば当該クラスタを構成する各文章に関する情報として、例えば当該クラスタに属する実文章や仮想代表文章の文章番号、さらにはそのステップ数などを大分類データ23として出力し記憶部20へ格納する(ステップ167)。そして、大分類ラベル生成手段73を用いて、新たなクラスタに含まれていた仮想代表文章について、その元となるクラスタに対して後述の大分類ラベル生成処理を行う(ステップ168)。

[0067] このようにして、ステップ161ー168までを1ステップとして、DTマトリクスに対する変形処理により新たなクラスタを生成するとともに、そのクラスタをその仮想代表文章で置換することにより新たなDTマトリクスを生成するクラスタリング処理を実行し、その後、ステップ161へ戻って、新たなDTマトリクスを用いたクラスタリング処理を繰り返し実行する。これにより、クラスタリング処理の繰り返しステップで生成されたクラスタには、

実文章だけでなく仮想代表文章すなわち他のクラスタも含まれることになり、各文章の大分類が得られることになる。

[0068] 図19に、大分類生成処理の実行例を示す。ここでは初期状態として、記憶部20の文章集合21に文章aーkが格納されているものとする。そして、1回目のクラスタリング処理であるステップS1で、文章a，bからクラスタ301が生成され、その仮想代表文章V1が生成されている。同様にして、文章c，dからクラスタ302が生成され、その仮想代表文章V2が生成されており、さらに文章e，fからクラスタ303が生成され、その仮想代表文章V3が生成されている。

[0069] これにより、ステップS1終了時点では、文章a，b，c，d，e，fがDTマトリクス上から削除され、文章gーkと仮想代表文章V1，V2，V3からなる新たなDTマトリクスを用いたステップS2が実行される。2回目のステップS2では、仮想代表文章V1と文章gからクラスタ304が生成され、その仮想代表文章V4が生成されている。この際、図18のステップ168における大分類ラベル生成処理では、クラスタ304に仮想代表文章V1が含まれていることから、その仮想代表文章V1の元となるクラスタ301に対する大分類ラベルが生成される。

[0070] ここで、図20を参照して、大分類ラベル生成処理について説明する。大分類ラベル生成手段73は、まず、大分類生成処理における現在のステップが、新たなクラスタが見つからなかった最終ステップかどうか判断する（ステップ170）。このとき、最終ステップでなければ（ステップ170：NO）、図18のステップ162で識別された新たなクラスタのうちから当該ラベル生成処理が未処理のクラスタを任意に1つ選択し（ステップ171）、そのクラスタに仮想代表文章が含まれているかどうか判断する（ステップ172）。なお、実文章と仮想代表文章とは、その文章番号などで識別すればよい。ここで、仮想代表文章が含まれている場合にのみ（ステップ172：YES）、DTマトリクス上でその仮想代表文章と強連結しているタームのキーワードから、その仮想代表文章の元のクラスタのラベルを生成する（ステップ173）。

[0071] そして、当該ラベル生成処理が未処理のクラスタがあれば（ステップ174：NO）、ステップ171に戻って未処理クラスタに対するラベル生成処理ステップ171ー173を繰り返し実行し、各クラスタに対する処理が終了した時点で（ステップ174：YES）、一連

の大分類生成処理を終了する。

[0072] また、ステップ170において、大分類生成処理における現在のステップが最終ステップであった場合は(ステップ170:YES)、その最終ステップの時点においてDTマトリクスを構成する各文章から、当該ラベル生成処理が未処理の仮想代表文章を任意に1つ選択し(ステップ180)、DTマトリクス上でその仮想代表文章と強連結しているタームのキーワードから、その仮想代表文章の元のクラスタのラベルを生成する(ステップ181)。そして、当該ラベル生成処理が未処理の仮想代表文章があれば(ステップ182:NO)、ステップ180に戻って未処理の仮想代表文章に対するラベル生成処理ステップ180, 181を繰り返し実行し、各仮想代表文章に対する処理が終了した時点で(ステップ182:YES)、一連の大分類生成処理を終了する。

[0073] したがって、図19のステップS2では、クラスタ304に仮想代表文章V1が含まれていることから、ステップS2の処理開始時点におけるDTマトリクス上でその仮想代表文章V1と強連結しているタームのキーワードから、その仮想代表文章V1の元のクラスタ301のラベルL1が生成される。以下、同様にして、ステップS3では、仮想代表文章V2と文章hからクラスタ305が生成され、その仮想代表文章V5が生成されている。そして、仮想代表文章V2の元のクラスタ305のラベルL2が生成される。

[0074] 次のステップS4では、仮想代表文章V4, V5と文章iからクラスタ306が生成されて、その仮想代表文章V6が生成されるとともに、仮想代表文章V3と文章jからクラスタ307が生成されて、その仮想代表文章V7が生成されている。そして、仮想代表文章V4の元のクラスタ304のラベルL4が生成されるとともに、仮想代表文章V5の元のクラスタ305のラベルL5が生成され、さらに仮想代表文章V3の元のクラスタ303のラベルL3が生成されている。続くステップS5では、仮想代表文章V6と文章kからクラスタ308が生成されて、その仮想代表文章V8が生成されている。そして、仮想代表文章V6の元のクラスタ306のラベルL6が生成されている。

[0075] 大分類生成手段71では、このようにしてクラスタリング処理(ステップ161ー168)を繰り返し実行し、図18のステップ163で新たなクラスタが見つからなかった場合は(ステップ163:NO)、最終ステップとして、大分類ラベルの付いていないクラスタに対する大分類ラベル生成処理を実行し(ステップ169)、一連の大分類生成処理を終了す

る。

[0076] これにより、図19の最終ステップでは、その時点のDTマトリクス上で、仮想代表文章V8と強連結しているタームのキーワードから、その仮想代表文章V8の元のクラスタ308のラベルL8が生成され、同様にして仮想代表V7の元のクラスタ307のラベルL7が生成される。

[0077] 図22に、初期状態におけるDTマトリクスの生成例を示す。各文章Di内にタームTjが存在する場合、列方向(横方向)に配置された文章Diと行方向(縦方向)に配置されたタームTjとの交点にドットが配置されており、タームTjが存在しない場合は空白となっている。なお、このDTマトリクスのうち、エリア310には実文章が横軸に配置されており、エリア311は仮想代表文章の配置用のため初期状態では空白となっている。図23に、最終ステップにおけるDTマトリクスの生成例を示す。この例では、大分類生成処理によりエリア310の実文章が削除されてほとんど空白となり、エリア311の仮想代表文章に置換されていることがわかる。

[0078] このように、本実施の形態では、DTマトリクスに対する変形処理により新たなクラスタを生成するとともに、そのクラスタをその仮想代表文章で置換することにより新たなDTマトリクスを生成するクラスタリング処理を繰り返し実行するようにしたので、新たなDTマトリクスから順次新たなクラスタ、すなわちクラスタを含むより大きなクラスタすなわち大分類が得られる。これにより、記憶部20の大分類データ23として、図19に示されているように、各文章a－kのみを要素とする分類、例えばクラスタ301－303だけでなく、1つ以上のクラスタを含むより大きな分類すなわち大分類として、クラスタ304－308が得られる。

[0079] さらに、上記クラスタリング処理をDTマトリクス上で新たなクラスタが識別されなくなるまで繰り返し実行するようにしたので、各文章からボトムアップ的に階層化クラスタリングが行われ、これらクラスタ301－308間すなわち大分類間の階層的関係をツリー構造として可視化することができる。

[0080] なお、以上では、大分類生成処理(図18参照)で、大分類ラベル生成処理(ステップ168, 169)を行う場合を例として説明したが、大分類ラベルが不要な場合は、大分類生成処理から大分類ラベル生成処理を省略してもよい。また、大分類ラベル生

成処理は、大分類生成処理と連携させて行う必要はなく、大分類生成処理が終了した後、必要に応じて大分類ラベル生成処理(図20参照)を独立して行ってもよい。

また、以上で説明した各実施の形態については、それぞれ個別に実施してもよく、両者を組み合わせて実施してもよい。

産業上の利用可能性

[0081] 本発明にかかる文章分類装置および方法は、各種内容の文書が含まれる文書集合を分類する場合に有用であり、特に、インターネットなどの通信網を介して不特定多数の利用者が入力したコメントやアンケート(自由文)などを分類分析するのに適している。

請求の範囲

[1] 1つ以上の単語からなるタームを複数有するタームリストと、

文章集合に含まれる各文章と前記各タームとの関係を2次元表現したDTマトリクスを生成するDTマトリクス生成手段と、

グラフ理論で用いられるDM分解法に基づいて前記DTマトリクス生成手段で得られたDTマトリクスを変形することにより、関連する文章のブロックからなるクラスタを有する変形DTマトリクスを生成するDTマトリクス変形手段と、

このDTマトリクス変形手段で得られた変形DTマトリクス上の各クラスタと、これらクラスタで分類される前記各文章との関係に基づき、前記文章集合に関する分類を生成する分類生成手段とを備えることを特徴とする文章分類装置。

[2] 請求項1に記載の文章分類装置において、

前記分類生成手段は、前記DTマトリクス変形手段で得られた変形DTマトリクス上のクラスタごとに、当該クラスタに属する文章を同一分類として出力する文章分類手段からなることを特徴とする文章分類装置。

[3] 請求項2に記載の文章分類装置において、

任意の前記クラスタに属する各文章と強連結をなす各タームを、当該クラスタの分類を示すラベルとして出力するラベル生成手段をさらに備えることを特徴とする文章分類装置。

[4] 請求項2に記載の文章分類装置において、

前記変形DTマトリクスでの文章の並び順序に応じて、任意の前記クラスタに属する文章またはすべての文章を順に出力する文章編成手段をさらに備えることを特徴とする文章分類装置。

[5] 請求項2に記載の文章分類装置において、

任意の前記文章を構成する各文のうち、当該文章と強連結をなすタームを含む文を、当該文章の要約として出力する要約作成手段をさらに備えることを特徴とする文章分類装置。

[6] 請求項2に記載の文章分類装置において、

前記タームリストに対して任意のタームを追加または削除するタームリスト編集手段

と、

このタームリスト編集手段による編集前後のタームリストを用いて前記DTマトリクス生成手段によりそれぞれDTマトリクスを生成し、これらDTマトリクスから当該編集の有用性を示す指標を生成して出力する指標生成手段とをさらに備えることを特徴とする文章分類装置。

[7] 請求項1に記載の文章分類装置において、

前記分類生成手段は、

変形DTマトリクス上のクラスタごとに、当該クラスタに属する各文章のタームから仮想代表文章を生成する仮想代表文章生成手段と、

前記DTマトリクス生成手段で生成したDTマトリクスを初期状態として、前記DTマトリクス変形手段でDTマトリクスから生成された変形DTマトリクス上のクラスタごとに前記仮想代表文章生成手段で仮想代表文章を生成し、前記仮想代表文章を当該変形DTマトリクスに追加するとともに前記仮想代表文章のクラスタに属する文章を当該変形DTマトリクスから削除して次のクラスタリング処理に用いる新たなDTマトリクスを生成し、前記クラスタごとに少なくとも当該クラスタを構成する文章に関する情報を大分類データとして出力する、というクラスタリング処理を繰り返し行うことにより前記文章の大分類を生成する大分類生成手段と

からなることを特徴とする文章分類装置。

[8] 請求項7に記載の文章分類装置において、

前記大分類生成手段は、前記クラスタリング処理で、前記変形DTマトリクスからクラスタが得られなくなった場合に、前記クラスタリング処理の繰り返しを終了することを特徴とする文章分類装置。

[9] 請求項7に記載の文章分類装置において、

前記クラスタリング処理で得られた各クラスタのうち、当該クラスタに仮想代表文章が含まれている場合は、その仮想代表文章と強連結をなすタームから、当該仮想代表文章の元のクラスタのラベルを生成する大分類ラベル生成手段をさらに備えることを特徴とする文章分類装置。

[10] 文章集合に含まれる各文章と1つ以上の単語からなるタームを複数有するタームリ

ストの各タームとの関係を2次元表現したDTマトリクスを生成するDTマトリクス生成ステップと、

グラフ理論で用いられるDM分解法に基づいて前記DTマトリクスを変形することにより、関連する文章のブロックからなるクラスタを有するクラスタを有する変形DTマトリクスを生成するDTマトリクス変形ステップと、

前記変形DTマトリクス上の各クラスタと、これらクラスタで分類される前記各文章との関係に基づき、前記文章集合に関する分類を生成する分類生成ステップと

を備えることを特徴とする文章分類方法。

[11] 請求項10に記載の文章分類方法において、

前記分類生成ステップは、前記変形DTマトリクス上のクラスタごとに、当該クラスタに属する文章を同一分類として出力する文章分類ステップからなることを特徴とする文章分類方法。

[12] 請求項11に記載の文章分類方法において、

任意の前記クラスタに属する各文章と強連結をなす各タームを、当該クラスタの分類を示すラベルとして出力するステップをさらに備えることを特徴とする文章分類方法。

[13] 請求項11に記載の文章分類方法において、

前記変形DTマトリクスでの文章の並び順序に応じて、任意の前記クラスタに属する文章またはすべての文章を順に出力するステップをさらに備えることを特徴とする文章分類方法。

[14] 請求項11に記載の文章分類方法において、

任意の前記文章を構成する各文のうち、当該文章と強連結をなすタームを含む文を、当該文章の要約として出力するステップをさらに備えることを特徴とする文章分類方法。

[15] 請求項11に記載の文章分類方法において、

前記タームリストに対して任意のタームを追加または削除するステップと、

編集前後のタームリストを用いてそれぞれDTマトリクスを生成し、これらDTマトリクスから当該編集の有用性を示す指標を生成して出力するステップとをさらに備えるこ

とを特徴とする文章分類方法。

[16] 請求項10に記載の文章分類方法において、

前記分類生成ステップは、

変形DTマトリクス上のクラスタごとに、当該クラスタに属する各文章のタームから仮想代表文章を生成する仮想代表文章生成ステップと、

前記DTマトリクス生成ステップで生成したDTマトリクスを初期状態として、前記DTマトリクス変形ステップでDTマトリクスから生成された変形DTマトリクス上のクラスタごとに前記仮想代表文章生成ステップで仮想代表文章を生成するステップと、前記仮想代表文章を当該変形DTマトリクスに追加するとともに前記仮想代表文章のクラスタに属する文章を当該変形DTマトリクスから削除して次のクラスタリング処理に用いる新たなDTマトリクスを生成するステップと、前記クラスタごとに少なくとも当該クラスタを構成する文章に関する情報を大分類データとして出力するステップと、からなるクラスタリング処理を繰り返し行うことにより前記文章の大分類を生成する大分類生成ステップと

からなることを特徴とする文章分類方法。

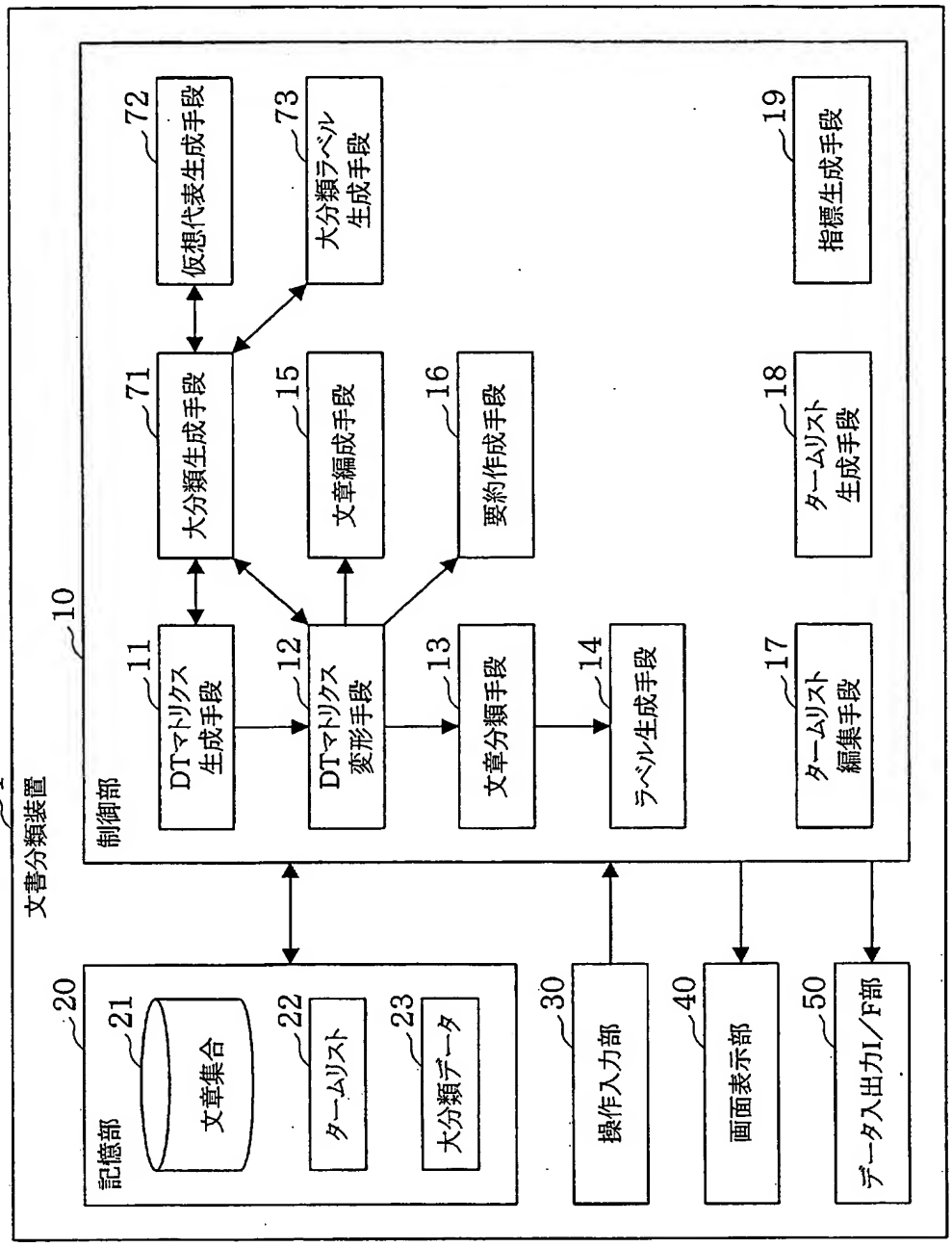
[17] 請求項16に記載の文章分類方法において、

前記大分類生成ステップは、前記クラスタリング処理で、前記変形DTマトリクスからクラスタが得られなくなった場合に、前記クラスタリング処理の繰り返しを終了することを特徴とする文章分類方法。

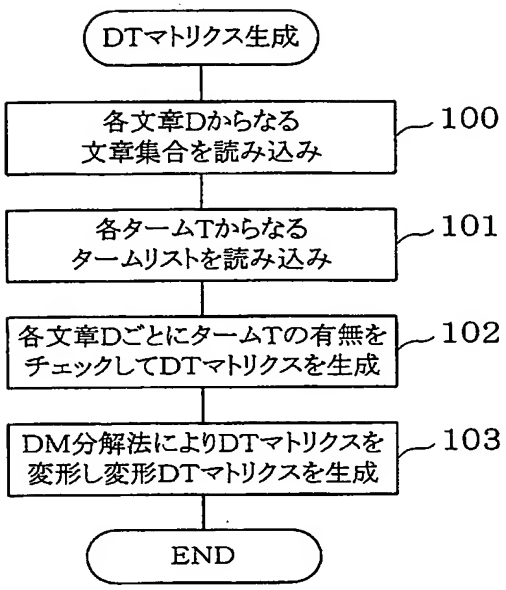
[18] 請求項16に記載の文章分類方法において、

前記クラスタリング処理で得られた各クラスタのうち、当該クラスタに仮想代表文章が含まれている場合は、その仮想代表文章と強連結をなすタームから、当該仮想代表文章の元のクラスタのラベルを生成する大分類ラベル生成ステップをさらに備えることを特徴とする文章分類方法。

[図1]


1
文書分類装置
10
制御部
11
DTマトリクス生成手段
12
DTマトリクス変形手段
13
文章分類手段
14
ラベル生成手段
15
文章編成手段
16
要約作成手段
17
タームリスト編集手段
18
タームリスト生成手段
19
指標生成手段
71
大分類生成手段
72
仮想代表生成手段
73
大分類ラベル生成手段
20
記憶部
21
文章集合
22
タームリスト
23
大分類データ
30
操作入力部
40
画面表示部
50
データ入出力I/F部

[図2]



[図3]

文章集合 21

Di	回答者	文章内容
1	W0039411	カラオケも楽しいですね。リズムにのって歌えるし間違っても誰も止めないですもんね。団でうたうのは周りの目も厳しいし頑張ら…
351	M0014787	現実逃避？なのかも知れませんが、私はストレスを感じたときは、ひたすら眠るようにしています。いい夢を見れば気分爽快なん…
289	M0010523	＞好きな香りの入浴剤を入れて、のんびり入浴。＞入浴しながら読書するとストレス解消になります。ぼくも本を持ち込んで読ん…
319	W0013732	マッサージって気持ちいいけど、ツボを心得ていない人にされるとかえってストレスになったりするのよね。マッサージも行きつけ…
57	W0039210	いいですね。気持ちよさそう!!! 私のように全然ばっとに当たらなければ、逆にストレスになってしまうかも… こういうストレスの…
72	W0039200	私もプールに通っています。ただがんばりすぎると逆に疲れるので三十分程度にして好きなように泳いだり歩いたりしています。…
111	W0012712	先日たった一泊ですが何年ぶりかで旅行に行ってきました。食事の準備や布団の上げ下ろしもしなくていいのがこんな幸せな…
337	W0013147	＞私のストレス解消法はなんといってもカーデニング！＞旦那も子供も文句ばかり言うけど、花は何の文句も言わずに咲いてく…
360	W0015958	3歳になる息子がいます。公園に遊びに行ったりしますが、大の大人ならブランコなんか恥ずかしくて乗れませんが、子供と一緒…
32	W0016759	私も不良主婦なんです。胸を張って家事は完璧ですなんて言えないのに、2～3ヶ月に1回くらいの割で夜遊びします。7時ごろ…
		⋮

[図4]

タームリスト 22

Tj	キーワード前		キーワード後		重要度
1	ストレス	名詞－一般	解消	名詞－サ変接続	256
2	解消	名詞－サ変接続	ストレス	名詞－一般	256
3	ストレス	名詞－一般	仕事	名詞－サ変接続	117
4	仕事	名詞－サ変接続	ストレス	名詞－一般	117
5	とき	名詞－非自立－副詞可能	ストレス	名詞－一般	116
6	吸う	動詞－自立	ストレス	名詞－一般	99
7	的	名詞－接尾－形容動詞語幹	ストレス	名詞－一般	88
8	行く	動詞－自立	ストレス	名詞－一般	86
9	寝酒	名詞－一般	晩酌	名詞－一般	85
10	晩酌	名詞－一般	寝酒	名詞－一般	85
11	人	名詞－接尾－助数詞	ストレス	名詞－一般	83
12	子供	名詞－一般	ストレス	名詞－一般	77
			⋮		

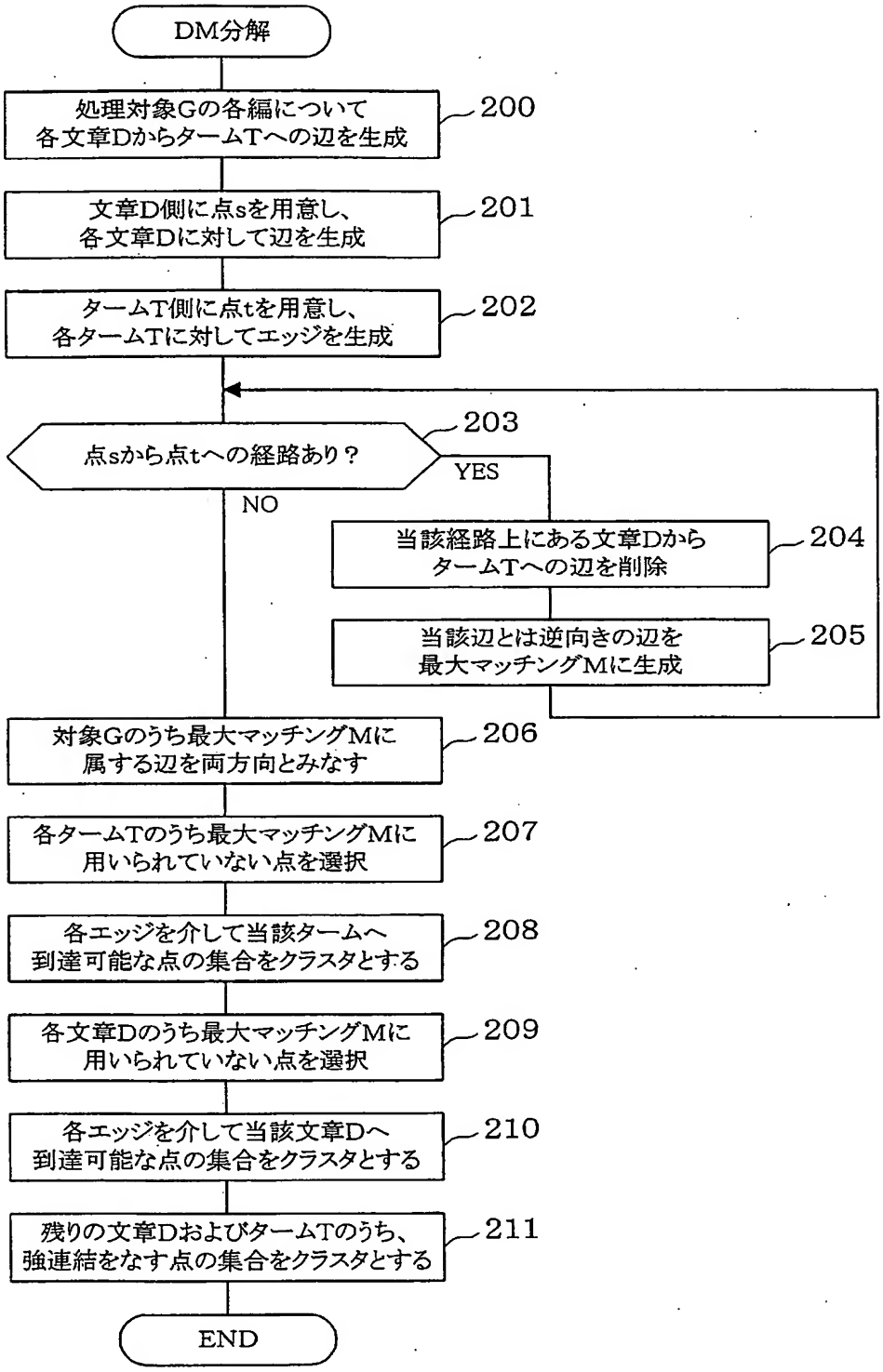
[図5]

DTマトリクス 11A

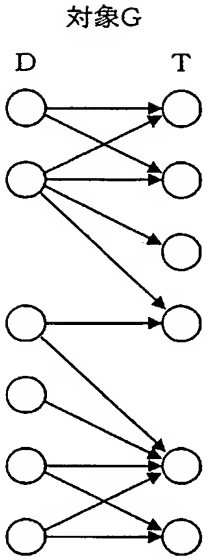
		文章D							
		D1	D2	D3	D4	D5	D6	⋯⋯	Dm
タームT	T1	0	1	0	0	0	0		0
	T2	0	1	0	1	0	0		0
	T3	0	0	0	0	0	1		0
	T4	1	0	0	0	0	0		0
	T5	0	0	0	0	0	0		1
	T6	0	0	0	0	1	0		1
	T7	1	0	0	0	0	0		0
	T8	0	0	0	0	0	0		0
	⋮	⋮		⋮				⋱	⋮
	Tn	0	0	0	1	0	0	⋯⋯	0

0＝文章Diにターム Tjが存在しない
1＝文章Diにターム Tjが存在する

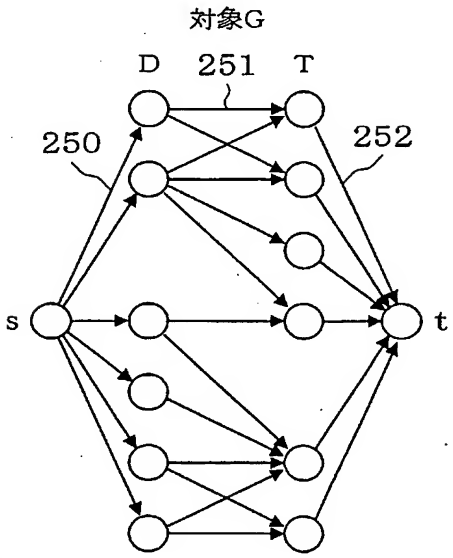
[図6]


DM分解
処理対象Gの各編について
各文章Dからターム Tへの辺を生成
200
文章D側に点sを用意し、
各文章Dに対して辺を生成
201
タームT側に点tを用意し、
各タームTに対してエッジを生成
202
点sから点tへの経路あり？
203
YES
NO
当該経路上にある文章Dから
タームTへの辺を削除
204
当該辺とは逆向きの辺を
最大マッチングMに生成
205
対象Gのうち最大マッチングMに
属する辺を両方向とみなす
206
各タームTのうち最大マッチングMに
用いられていない点を選択
207
各エッジを介して当該タームへ
到達可能な点の集合をクラスタとする
208
各文章Dのうち最大マッチングMに
用いられていない点を選択
209
各エッジを介して当該文章Dへ
到達可能な点の集合をクラスタとする
210
残りの文章DおよびタームTのうち、
強連結をなす点の集合をクラスタとする
211
END

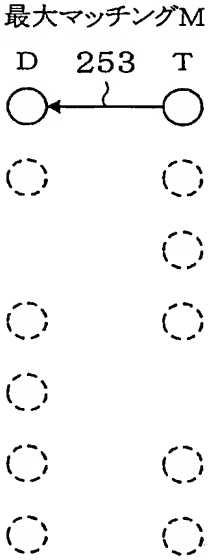
[図7A]



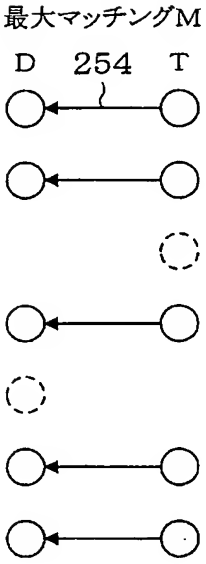
[図7B]



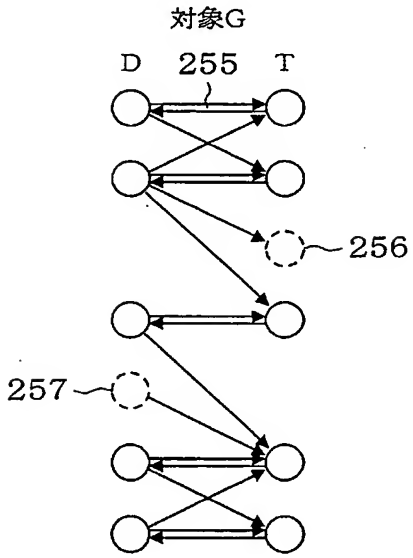
[図7C]



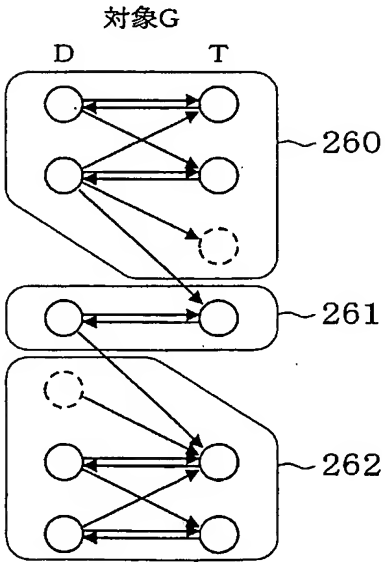
[図7D]



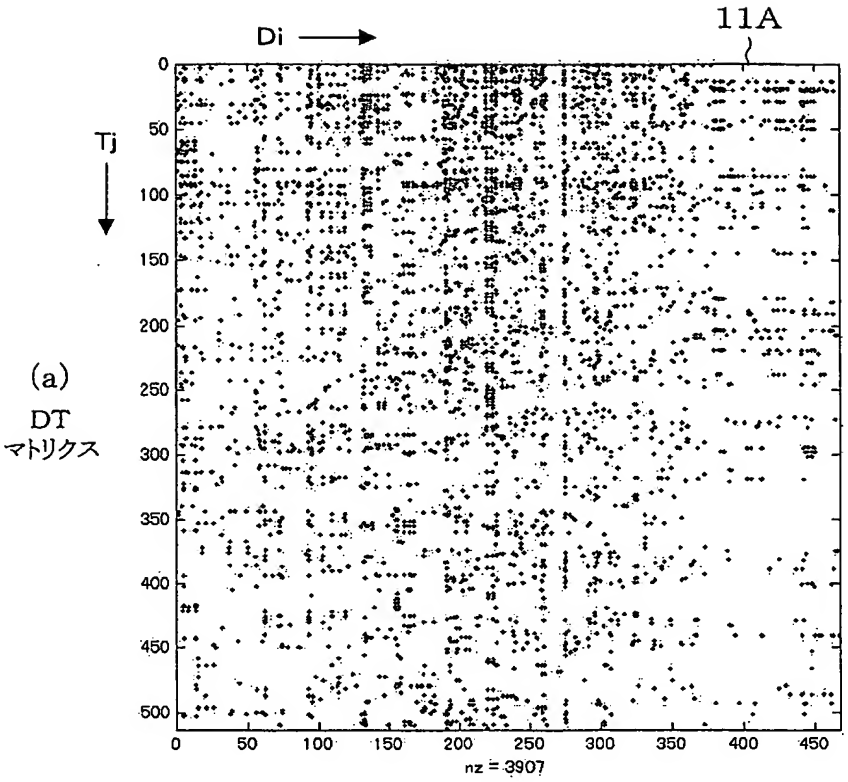
[図7E]



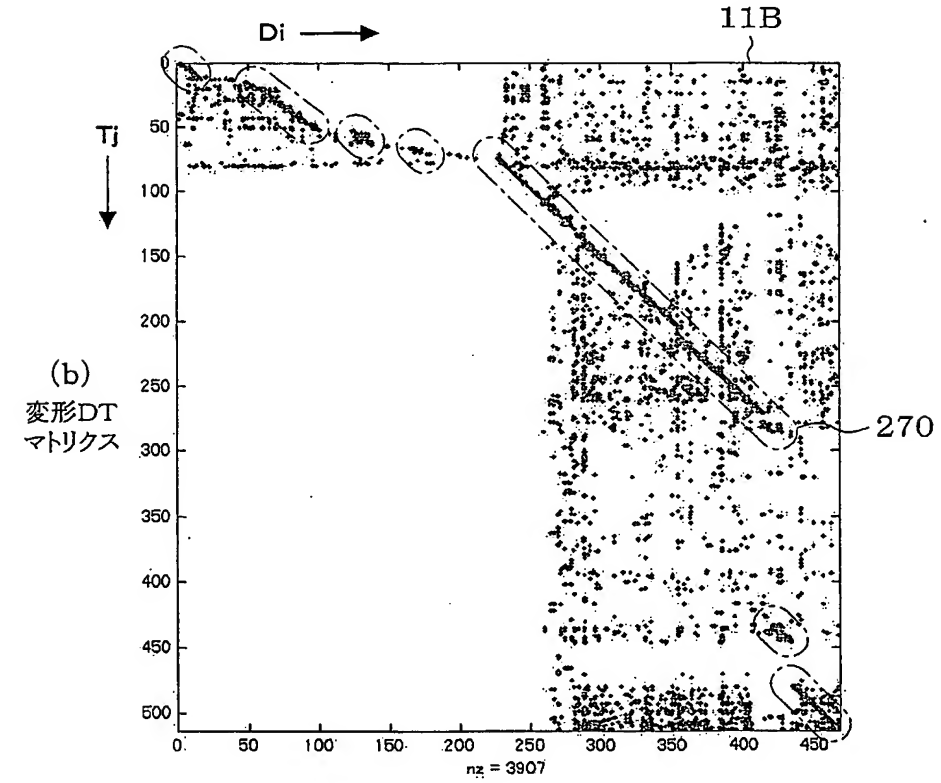
[図7F]



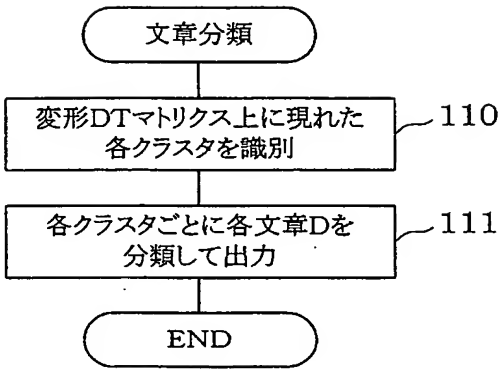
[図8A]



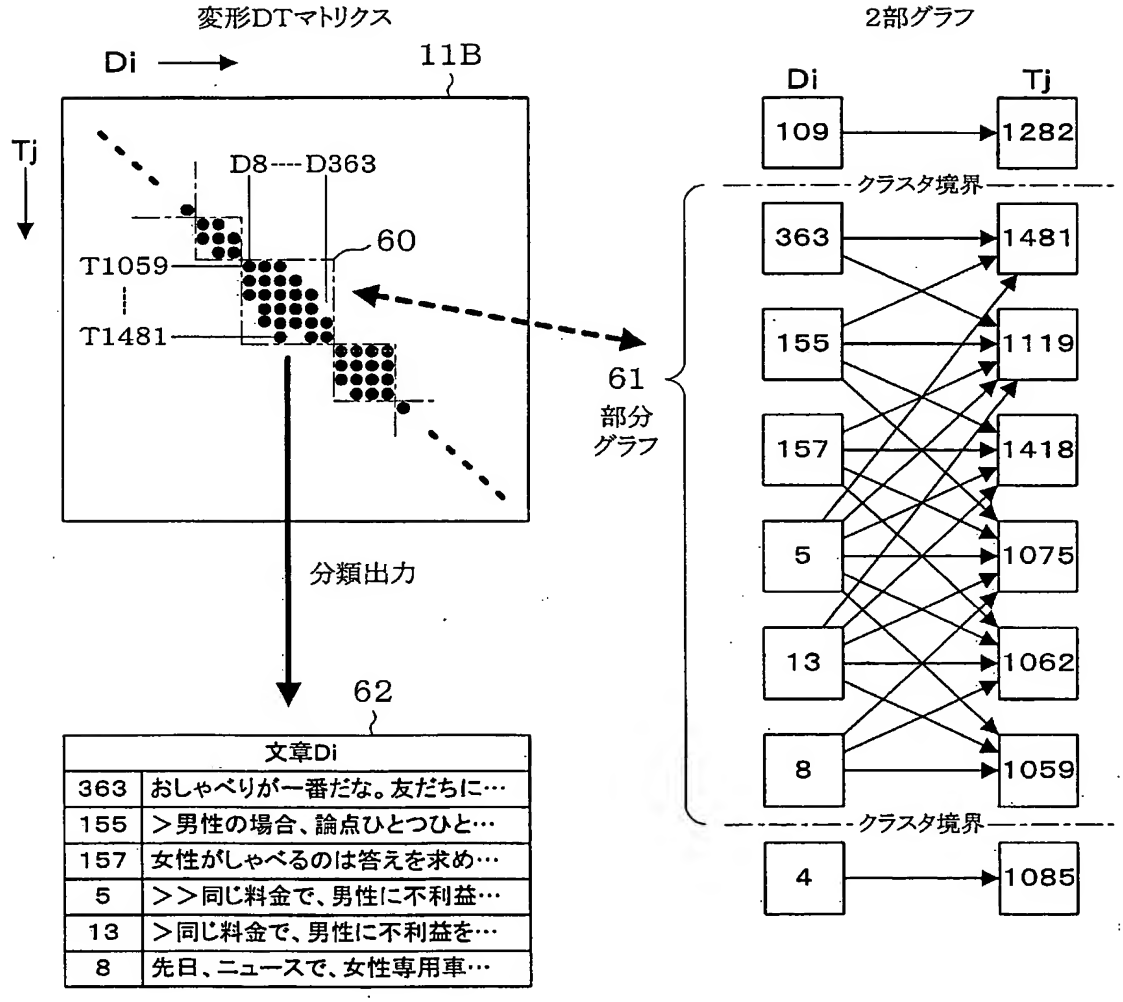
[図8B]



[図9]

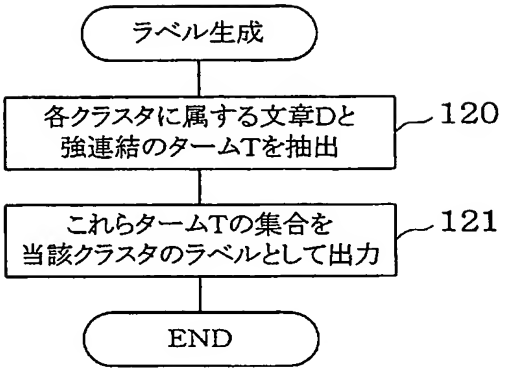


[図10]

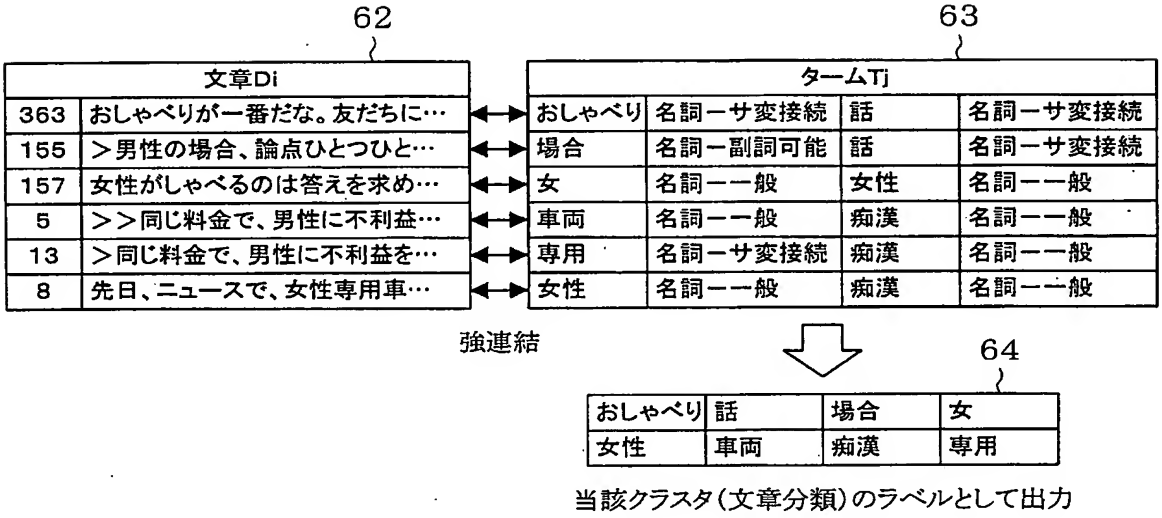


文章Di	
363	おしゃべりが一番だな。友だちに…
155	＞男性の場合、論点ひとつひと…
157	女性がしゃべるのは答えを求め…
5	＞＞同じ料金で、男性に不利益…
13	＞同じ料金で、男性に不利益を…
8	先日、ニュースで、女性専用車…

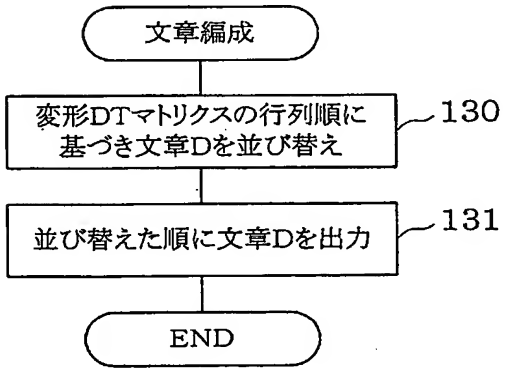
[図11]


ラベル生成
各クラスタに属する文章Dと
強連結のタームTを抽出
120
これらタームTの集合を
当該クラスタのラベルとして出力
121
END

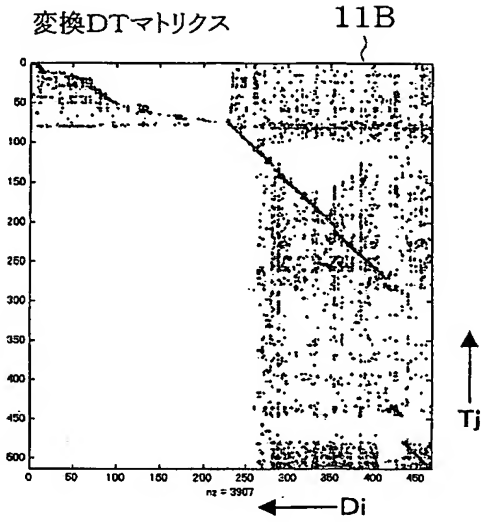
[図12]


62
63
文章Di
タームTj
363 おしゃべりが一番だな。友だちに…
155 ＞男性の場合、論点ひとつひと…
157 女性がしゃべるのは答えを求め…
5 ＞＞同じ料金で、男性に不利益…
13 ＞同じ料金で、男性に不利益を…
8 先日、ニュースで、女性専用車…
おしゃべり 名詞ーサ変接続 話 名詞ーサ変接続
場合 名詞ー副詞可能 話 名詞ーサ変接続
女 名詞ー一般 女性 名詞ー一般
車両 名詞ー一般 痴漢 名詞ー一般
専用 名詞ーサ変接続 痴漢 名詞ー一般
女性 名詞ー一般 痴漢 名詞ー一般
強連結
64
おしゃべり 話 場合 女
女性 車両 痴漢 専用
当該クラスタ(文章分類)のラベルとして出力

[図13]


文章編成
変形DTマトリクスの行列順に
基づき文章Dを並び替え
130
並び替えた順に文章Dを出力
131
END

[図14]



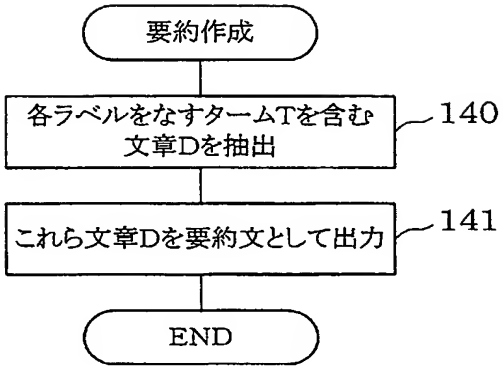
線形連立方程式の解が
求まる順序として、
変形DTマトリクス上に
半順序Gi≦Gjが存在

変形DTマトリクスにおける文章Dの並びに基づき各文書Dを並び替え

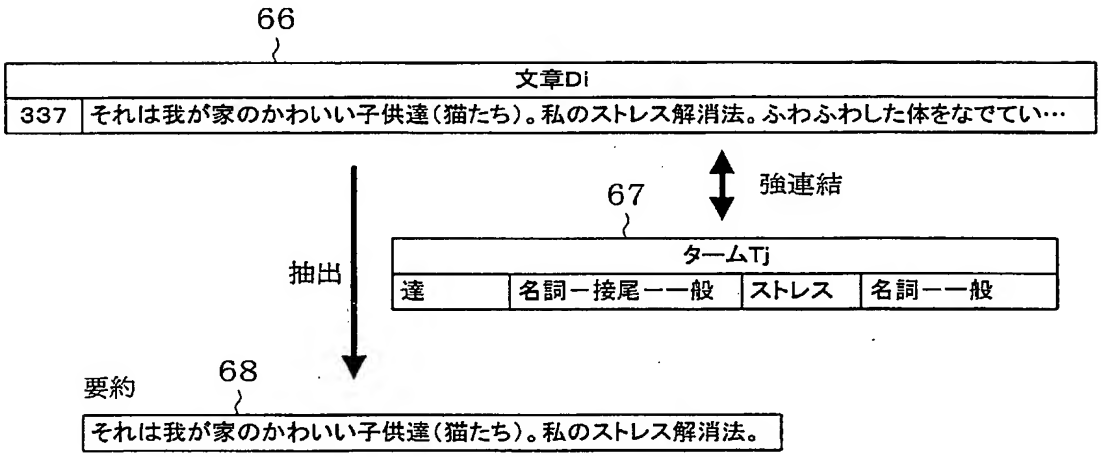
65

	文章Di
1	私はあんまりストレスを感じませんが… みんなそれぞれ違いますしストレスを感じてしまう…
351	現実逃避？なのかもれませんが、私はストレスを感じたときはひたすら眠るようにしています。
289	ぼくも本を持ち込んで読んだことがあるんですが、あがったときに本がブヨブヨ(笑)なんとか…
319	(笑)。マッサージって気持ちいいけど、ツボを心得ていない人にされるとかえってストレスに…
57	気持ちよさそう!!! 私のように全然バットに当たらなければ、逆にストレスになってしまうかも…
72	ただがんばりすぎると逆に疲れるので三十分程度にして好きなように泳いだり歩いたりして…
111	温泉にゆっくり入ってストレスも吹っ飛んでしまいました。
337	＞私のストレス解消法はなんといってもガーデニング！＞旦那も子供も文句ばっかり言うけ…
360	公園に遊びに行ったりしますが 大の大人なら ブランコなんか恥ずかしくて乗れませんが…
32	メッチャ楽しい！メンバーは子供会活動を一緒にやった人達8名。
390	子供の頃、銭湯に行ったことがあったけど大きくて泳げたりしたのがたのしかったです。
227	雑誌なんかを読みながら＞トリートメントをしたり、ボディマッサージなんかして＞なんか自…
	⋮

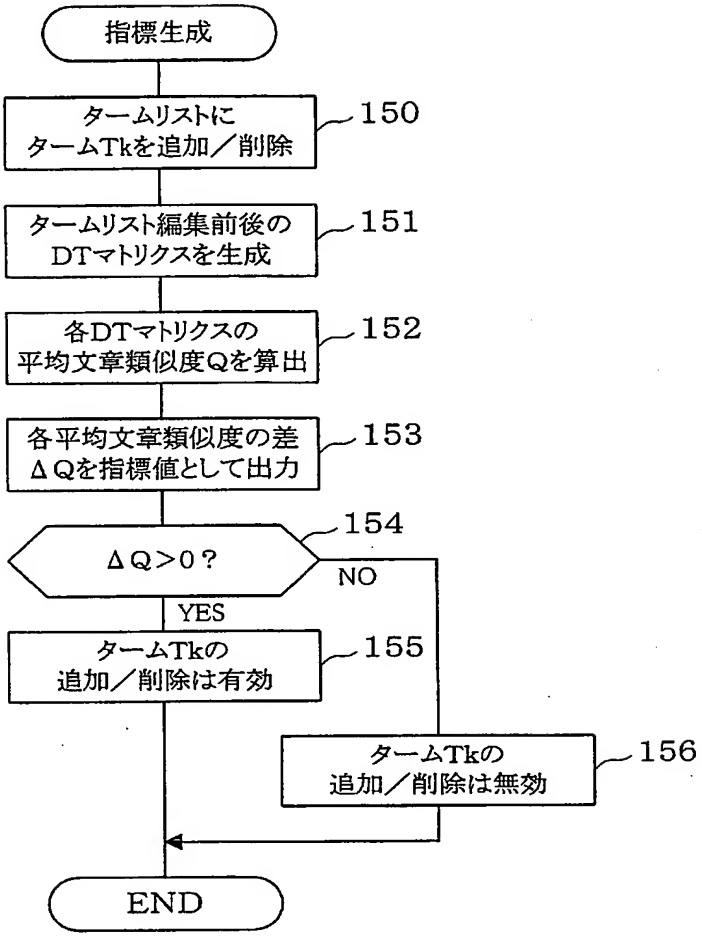
[図15]


要約作成
各ラベルをなすタームTを含む文章Dを抽出
140
これら文章Dを要約文として出力
141
END

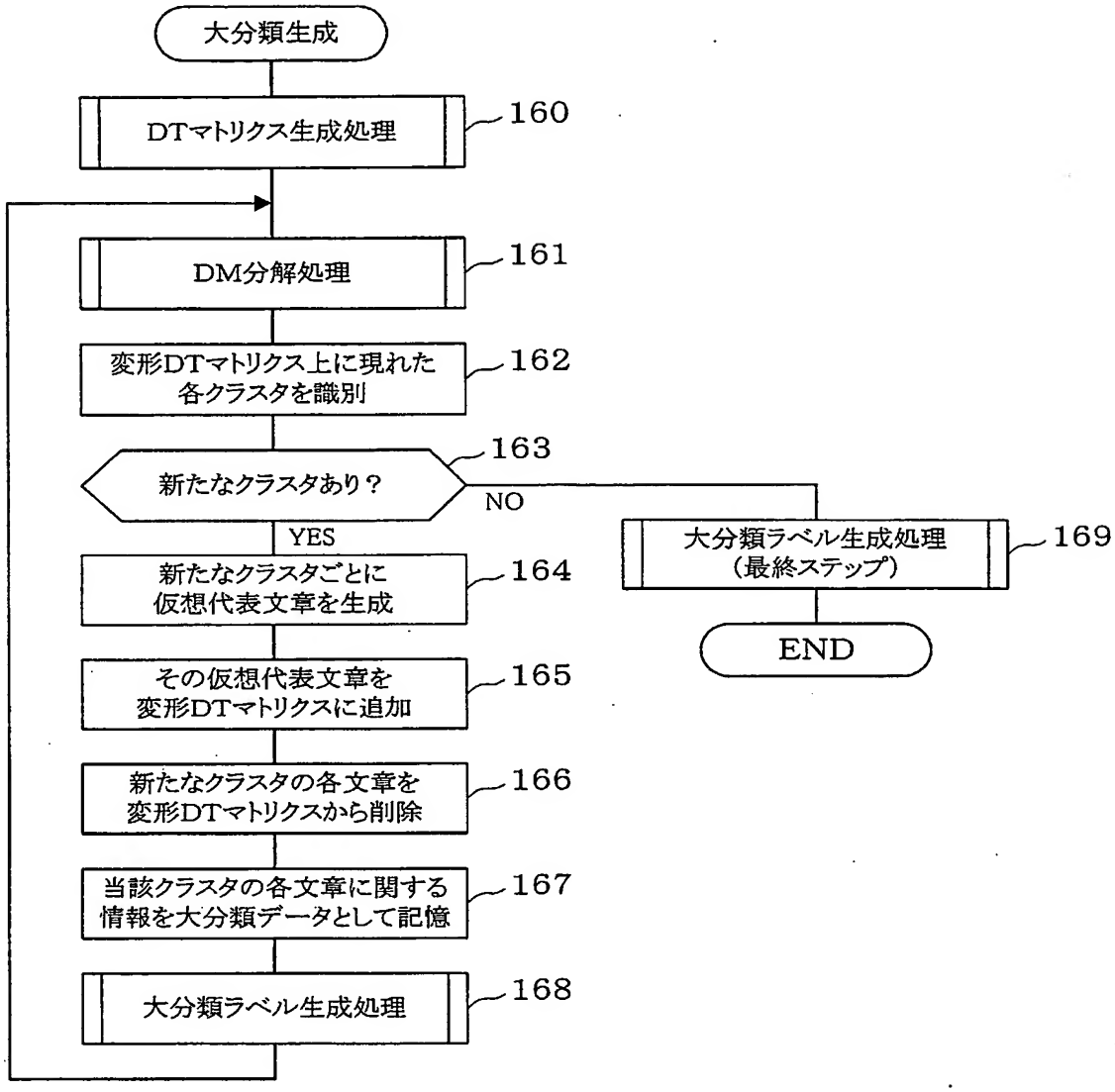
[図16]


66
文章Di
337
それは我が家のかわいい子供達(猫たち)。私のストレス解消法。ふわふわした体をなでてい…
強連結
67
タームTj
達
名詞ー接尾ー一般
ストレス
名詞ー一般
抽出
要約
68
それは我が家のかわいい子供達(猫たち)。私のストレス解消法。

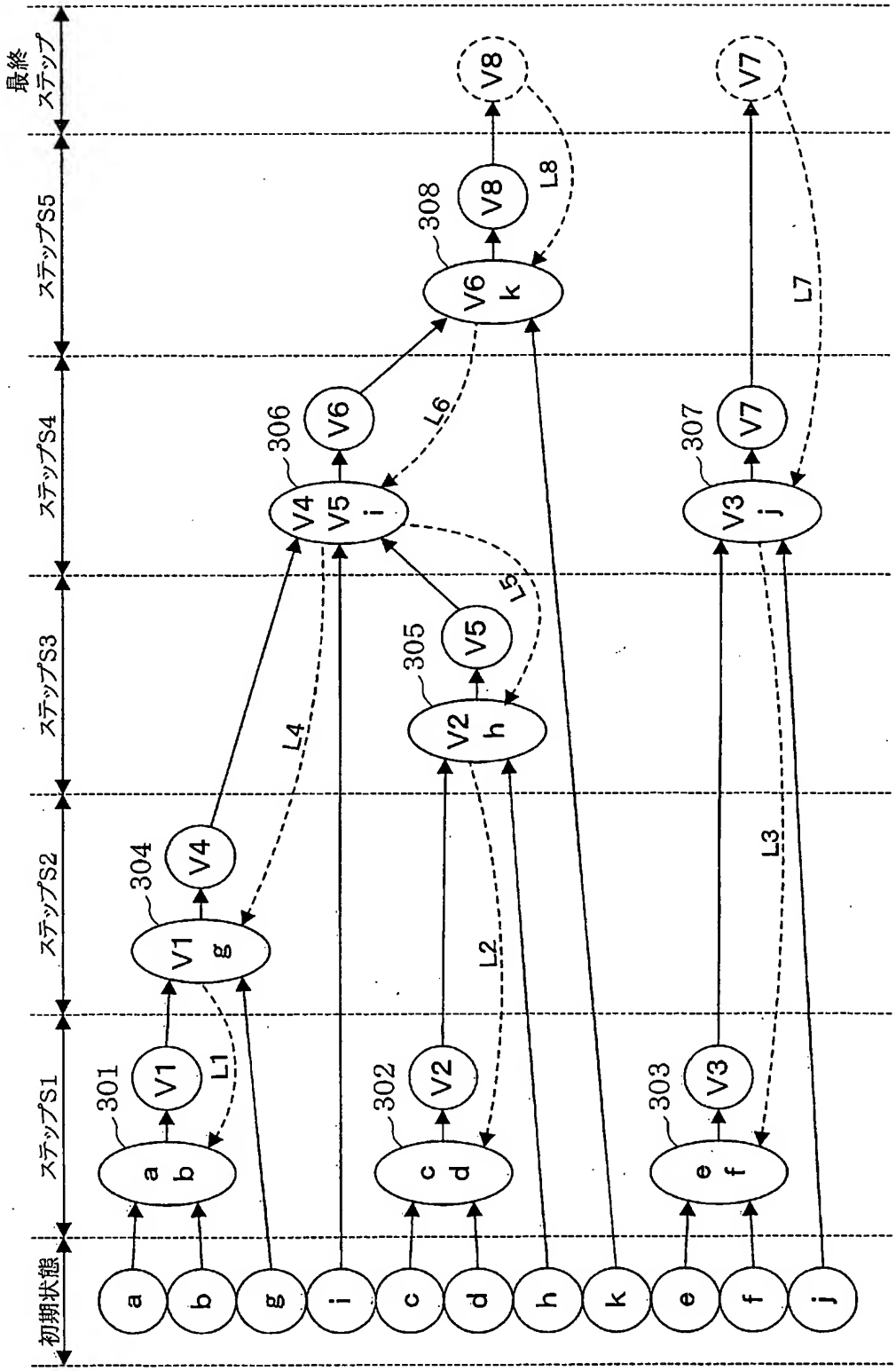
[図17]


指標生成
タームリストに
タームTkを追加／削除
150
タームリスト編集前後の
DTマトリクスを生成
151
各DTマトリクスの
平均文章類似度Qを算出
152
各平均文章類似度の差
ΔQを指標値として出力
153
ΔQ>0？
154
NO
YES
タームTkの
追加／削除は有効
155
タームTkの
追加／削除は無効
156
END

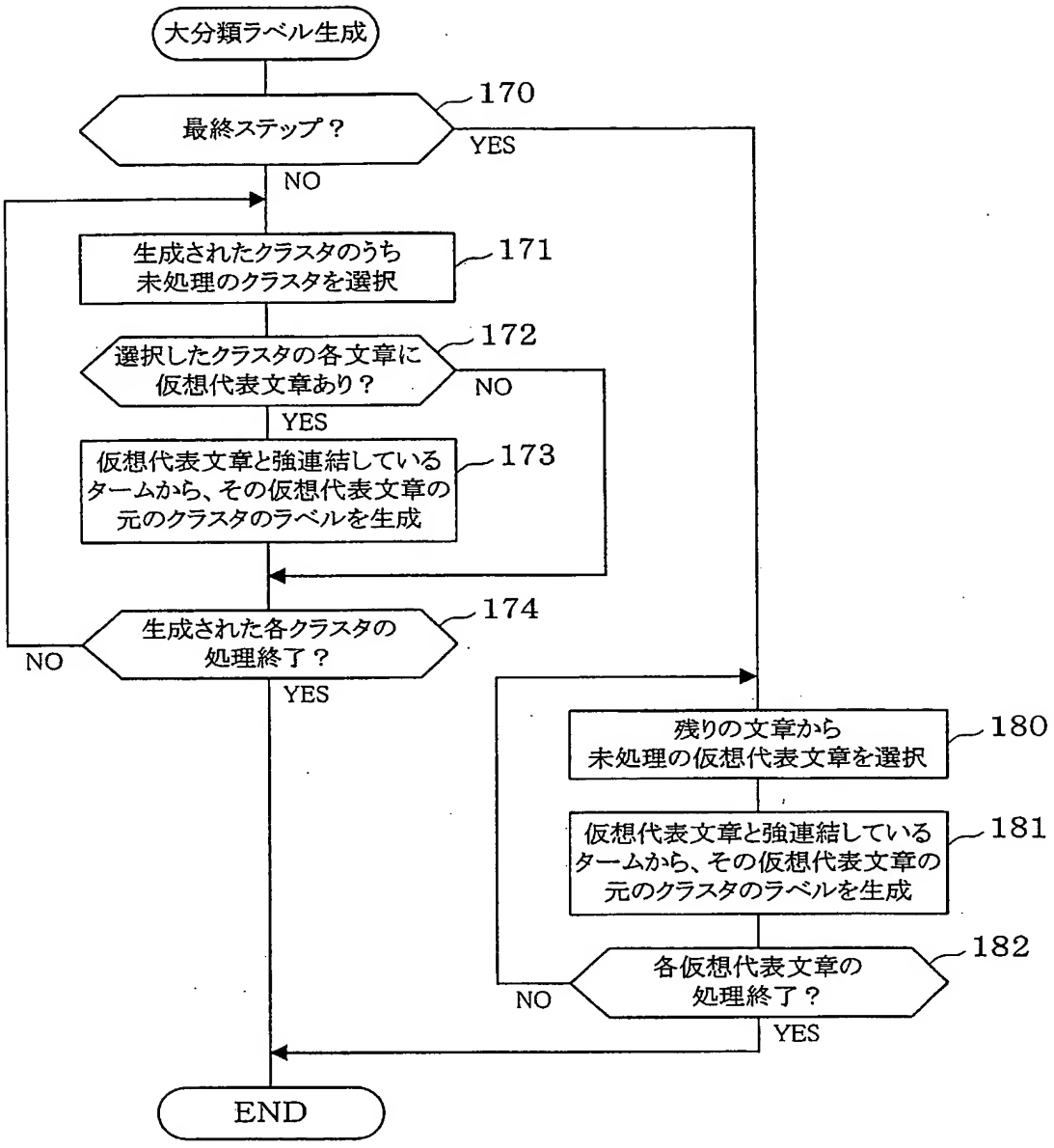
[図18]


大分類生成
DTマトリクス生成処理
160
DM分解処理
161
変形DTマトリクス上に現れた
各クラスタを識別
162
新たなクラスタあり？
163
NO
YES
大分類ラベル生成処理
（最終ステップ）
169
END
新たなクラスタごとに
仮想代表文章を生成
164
その仮想代表文章を
変形DTマトリクスに追加
165
新たなクラスタの各文章を
変形DTマトリクスから削除
166
当該クラスタの各文章に関する
情報を大分類データとして記憶
167
大分類ラベル生成処理
168

[図19]

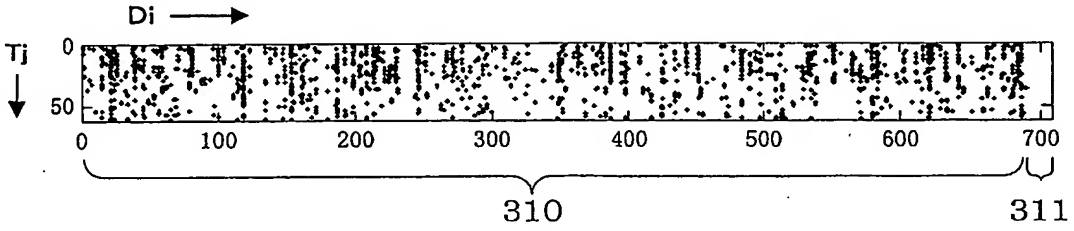

最終ステップ
ステップS5
ステップS4
ステップS3
ステップS2
ステップS1
初期状態
V8
V7
V6
V5
V4
V3
V2
V1
a
b
g
i
c
d
h
k
e
f
j
L1
L2
L3
L4
L5
L6
L7
L8
301
302
303
304
305
306
307
308

[図20]


大分類ラベル生成
最終ステップ？
170
YES
NO
生成されたクラスタのうち
未処理のクラスタを選択
171
選択したクラスタの各文章に
仮想代表文章あり？
172
NO
YES
仮想代表文章と強連結している
タームから、その仮想代表文章の
元のクラスタのラベルを生成
173
生成された各クラスタの
処理終了？
174
NO
YES
残りの文章から
未処理の仮想代表文章を選択
180
仮想代表文章と強連結している
タームから、その仮想代表文章の
元のクラスタのラベルを生成
181
各仮想代表文章の
処理終了？
182
NO
YES
END

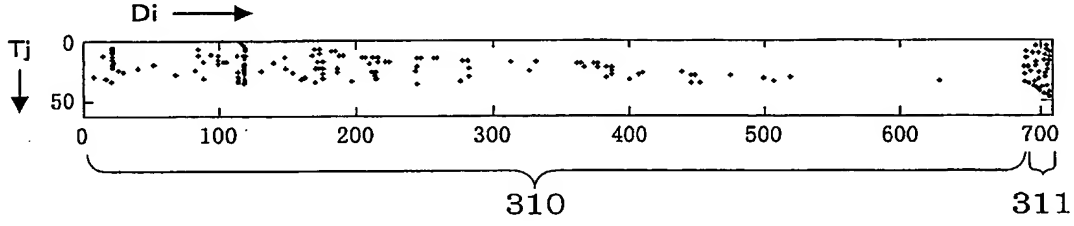
[図21]

DTマトリクス(初期状態)



[図22]

DTマトリクス(最終ステップ)



INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP2004/009341

A. CLASSIFICATION OF SUBJECT MATTER
Int.Cl[7] G06F17/30

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
Int.Cl[7] G06F17/30

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
Jitsuyo Shinan Koho 1922-1996 Toroku Jitsuyo Shinan Koho 1994-2004
Kokai Jitsuyo Shinan Koho 1971-2004 Jitsuyo Shinan Toroku Koho 1996-2004

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
JSTPlus(JOIS), WPI, INSPEC(DIALOG)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	Seiji HOTTA et al., "Omomi Tsuki Graph kara no Fuzzy Cluster Chushutsu", The Transactions of the Institute of Electronics, Information and Communication Engineers, 01 March, 2001 (01.03.01), Vol.J84-A, No.3, pages 351 to 359	1-18
Y	Yoichi SATO et al., "Tangokan Imi Kankei no Graph Rironteki Kaiseki", The Institute of Electronics, Information and Communication Engineers Gijutsu Kenkyu Hokoku, 18 March, 1991 (18.03.91), Vol.90, No.482, pages 25 to 32	1-18
Y	JP 2002-108894 A (Ricoh Co., Ltd.), 12 April, 2002 (12.04.02), Par. Nos. [0074], [0075], [0105] to [0113] (Family: none)	7-9,16-18

☒ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
12 August, 2004 (12.08.04)	31 August, 2004 (31.08.04)
Name and mailing address of the ISA/ Japanese Patent Office	Authorized officer
Facsimile No.	Telephone No.

Form PCT/ISA/210 (second sheet) (January 2004)

INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP2004/009341

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
P,X	Eiji MURAKAMI et al., "Hito no Kyoji ni yoru Kanren Chishiki Kakutoku Hoho", Dai 64 Kai Chishiki Base System Kenkyukai Shiryo (SIG-KBS-A304), The Japanese Society for Artificial Intelligence, 01 March, 2004 (01.03.04), pages 207 to 212	1-18

国際調査報告

国際出願番号　PCT／JP2004／009341

A.　発明の属する分野の分類（国際特許分類（IPC））

Int. Cl⁷　G06F17／30

B.　調査を行った分野

調査を行った最小限資料（国際特許分類（IPC））

Int. Cl⁷　G06F17／30

最小限資料以外の資料で調査を行った分野に含まれるもの

日本国実用新案公報	1922－1996年
日本国公開実用新案公報	1971－2004年
日本国登録実用新案公報	1994－2004年
日本国実用新案登録公報	1996－2004年

国際調査で使用した電子データベース（データベースの名称、調査に使用した用語）

JSTPlus（JOIS），WPI，INSPEC（DIALOG）

C.　関連すると認められる文献

引用文献のカテゴリー＊	引用文献名　及び一部の箇所が関連するときは、その関連する箇所の表示	関連する請求の範囲の番号
Y	堀田政二 他，重み付きグラフからのファジークラスタ抽出，電子情報通信学会論文誌，2001.03.01，Vol.J84-A，No.3，pp.351-359	1-18
Y	佐藤洋一 他，単語間意味関係のグラフ理論的解析，電子情報通信学会技術研究報告，1991.03.18，Vol.90，No.482，pp.25-32	1-18
Y	JP 2002-108894 A（株式会社リコー）2002.04.12 第74,75段落，第105段落－第113段落　（ファミリーなし）	7-9，16-18

☒ C欄の続きにも文献が列挙されている。　☐ パテントファミリーに関する別紙を参照。

＊　引用文献のカテゴリー

「A」特に関連のある文献ではなく、一般的技術水準を示すもの
「E」国際出願日前の出願または特許であるが、国際出願日以後に公表されたもの
「L」優先権主張に疑義を提起する文献又は他の文献の発行日若しくは他の特別な理由を確立するために引用する文献（理由を付す）
「O」口頭による開示、使用、展示等に言及する文献
「P」国際出願日前で、かつ優先権の主張の基礎となる出願の日の後に公表された文献
「T」国際出願日又は優先日後に公表された文献であって出願と矛盾するものではなく、発明の原理又は理論の理解のために引用するもの
「X」特に関連のある文献であって、当該文献のみで発明の新規性又は進歩性がないと考えられるもの
「Y」特に関連のある文献であって、当該文献と他の1以上の文献との、当業者にとって自明である組合せによって進歩性がないと考えられるもの
「&」同一パテントファミリー文献

国際調査を完了した日 12.08.2004	国際調査報告の発送日 31.8.2004
国際調査機関の名称及びあて先 日本国特許庁（ISA／JP） 郵便番号100－8915 東京都千代田区霞が関三丁目4番3号	特許庁審査官（権限のある職員） 高瀬　勤　　5M　3364 電話番号　03－3581－1101　内線　3597

様式PCT／ISA／210（第2ページ）（2004年1月）

C（続き）．　関連すると認められる文献

引用文献のカテゴリー＊	引用文献名　及び一部の箇所が関連するときは、その関連する箇所の表示	関連する請求の範囲の番号
P，X	村上英治 他，人の教示による関連知識獲得方法，第６４回知識ベースシステム研究会資料(SIG-KBS-A304)，社団法人人工知能学会，2004.03.01，pp.207-212	1-18